Evaluation of Estimates of Extreme Fatigue Load

Enhanced by Data from Questionnaires

Martin Olofsson

Abstract

When estimating parameters of the probability distribution of a random variable, there are sometimes more observations available on related auxiliary variables. The usefulness of these observations is evaluated for the case when a jointly normal distribution is assumed. Focus is put on estimates of extreme distribution quantiles and the derivation of their estimation accuracy. A linear regression model is applied to make use of the auxiliary information. In order to decrease the mean square error of the 95*th* percentile estimator by 25% compared to the case when the auxiliary information is not used, it is found that a coefficient of determination, of the regression model, of about 50% is required. Moreover, it is shown how the maximum number of auxiliary variables allowed in the model decreases with smaller sizes of the primary sample. One application of this result arises in the investigation of severe service fatigue load, for instance on an automobile. In addition to a limited number of thorough service fatigue load measurements, a questionnaire survey can be used to measure fatigue load-related customer characteristics on a much larger population sample.

Keywords: Customer correlation, fatigue load, fatigue test, reliability, usage enquiry, multivariate analysis, regression, inverted Wishart, auxiliary data.

MSC2000: 62G32, 62J05, 62H20, 62P30.

Acknowledgements

First of all I would like to thank Prof. Jacques de Maré, my research advisor, for all the support and encouragement he has given me. Besides, without Jacques' strong interest in engineering applications and, in particular, material fatigue, I would not have had the fortune to expose myself to this very special environment I think only a mathematical university department can offer. I thank each of you for being indispensable elements of this extraordinary atmosphere and also for showing tolerance towards a somewhat mathematically illiterate engineer (me, that is).

I would also like to mention my appreciation for the pleasant two month-visit at the company LMS Durability Technologies, in Kaiserslautern, Germany. There, Dr. Bernhart Gründer and his staff explained to me the controversial idea about fatigue load measurement with a questionnaire.

Finally, I also feel gratitude towards the National Network in Applied Mathematics and its director Dr. Uno Nävert. Uno has been responsible for my private monetary fortune the last couple of years.

Martin Olofsson
Göteborg, September 2000

Table of contents

1 Introduction 1

2 Customer correlation overview 5
- 2.1 Customer fatigue load measurements 6
- 2.2 Reference fatigue load measurement 8
- 2.3 Statistical modelling and inference 9
- 2.4 Fatigue load specification 10

3 Problem statement 11

4 Regression on a scalar, normal auxiliary variable 13
- 4.1 Estimation of response mean 14
- 4.2 Estimation of response variance 16
- 4.3 Estimation of distribution quantile 19
- 4.4 Alternative regression parameter estimators 24

5 Regression on a multivariate, normal auxiliary variable 27
- 5.1 Estimation of response mean 28
- 5.2 Estimation of response variance 30
- 5.3 Estimation of distribution quantile 35

6 Preliminaries on general auxiliary variable distribution 39

7 Summary and conclusions 41

References 43

1 Introduction

The time allowed for the manufacturing industry to develop new products is decreasing continuously, as the competition gets tougher. It is important to find the optimal design without the need of extra time and cost for redesigns. In material fatigue engineering, component strength optimization is sometimes exaggerated, due to insufficient knowledge about the service loading conditions and, consequently, high safety factors. Many of the needless design modifications can be avoided, if the fatigue load in the real customer environment is well known and the load design requirement is correctly put. Unfortunately, it is often a demanding and expensive operation to acquire enough information about the service loading, in particular when there exists a great individual variation in product usage. Also, even with good knowledge about the complex fatigue loading in service, the specification of a load design requirement is a difficult task.

A natural way to acquire necessary information about the customer fatigue load environment is to make field measurements of forces, strains and other operation-descriptive entities, during a reasonably long time of customer usage. However, even though onboard data loggers are becoming less and less expensive and, at the same time, more capable in terms of data storage, it may still not be possible to make measurements at a large enough number of customers, necessary for good population distribution estimates of loading characteristics.

In the automotive industry, as an example, most traditional fatigue design and test procedures are based on intelligent guesswork by experienced engineers. Durability test tracks have been developed to replicate some worst-case loading conditions. Tests made on these test tracks have then eventually become standardized verifying tests, by experience from many years of continuous use together with feedback from the service departments. As a consequence, fatigue load specifications often originate from measurements made on these test tracks. The value of experience must not be underestimated here, but it is a fact that this procedure often yields design loads which have very little to do with the actual customer fatigue load environment. Hence, a high safety factor is embedded and, consequently, some components become unnecessarily overdesigned. Another issue worth bearing in mind is how the customer environment and usage change over time and, hence, the need for updating of durability tests and design requirements.

In the most recent years many automotive manufacturers have made an effort to make their fatigue load design specifications more related to the actual customer environment. See for instance Thomas et al. (1999) and Wang et al. (1999). Methodologies with names such as “customer correlation” are becoming more and more well known. The aim of these methodologies is to estimate the fatigue load environment corresponding to a certain time of usage or mileage at, typically, a 95th percentile (severe) customer.

In this paper the customer correlation notion refers to one specific approach, for the specification of service fatigue load environment, very much influenced by ideas from people at the German company LMS Durability Technologies (http://www.lmsintl.com). This approach involves an appealing method to improve confidence, when estimating population distributions of fatigue load measures from service load measurements. The method relies on the existence of auxiliary customer or usage characteristics, which are possible to measure more extensively on the customer population compared to the primary fatigue load measures studied. A questionnaire survey is used for the measurement of the auxiliary data. The critical requirement on the auxiliary data is the need for a strong relation to the primary load measures.

A direct field measurement of forces, strains, etc., which later is to be post-processed and reduced to the fatigue load measures already mentioned, is, of course, an inevitable and important part of the method. However, for cost reasons, this time- and instrumentation-demanding measurement can only be made on a limited number of customers. The confidence improvement of the load distribution estimates is accomplished by a questionnaire survey, for the measurement of fatigue-related auxiliary data on the same small customer sample as well as a great addition of customers (if not the full population). A regression analysis is the statistical tool for the assessment of the auxiliary data. The use of a questionnaire, to measure auxiliary data such as individual product usage and operator characteristics, is the distinguishing property of this particular customer correlation approach.

One crucial step in the statistical analysis is to model the relation between some well-chosen fatigue load measures and the auxiliary data from the questionnaires. This analysis is based on the questionnaire answers from, and field measurements on, the small customer sample. When the most suitable regression model is found and its parameters estimated, it is used to predict the primary fatigue load measures also for customers from which only questionnaire answers are available.

Another important issue, which will be left unresolved here, is how the limited customer sample will be drawn. A stratification approach would probably add further efficiency to the method.

The aim of this work has been to investigate conditions necessary for improved fatigue load distribution estimates, regarding model size, coefficient of determination, size of the small customer sample, etc. The performance of the method is assessed as the confidence improvement over the case when the auxiliary questionnaire data are not used. If such a confidence improvement were to prove negligible, the method would not be very useful, of course (or the questionnaire design would have been inadequate).

In this first analysis, the method is evaluated for the simplified case when the auxiliary data are assumed to have the normal distribution, even though in practice this would be far from the truth. Some useful auxiliary characteristics may, for instance, be categorical. The method performance for the case with normal-distributed auxiliary variables will still be of interest, however. We also limit the investigation to only one unspecified fatigue load measure. The choice of suitable ways to reduce field data to a manageable number of fatigue load measures is outside the scope of this paper. However, some general principles are given in Section 2, "Customer correlation overview". Further, the fatigue load variable is assumed to be linearly dependent on the auxiliary data, with a normal-distributed random error added. Hence, the normal distribution is also inherited by the fatigue load variable itself (as the dependent variable).

In addition to the issues already mentioned, there are many interesting problems left for future studies. The already mentioned sampling strategy, when the small sample is drawn, probably has a great influence on the efficiency of the method. Also, before a conclusive evaluation of the method is possible, the effect of different auxiliary characteristic distributions, other than the normal, should be investigated. Finally, since in practice there exists more than one dimension in the intricate fatigue load space, a multivariate regression analysis on several load measures, possibly with correlated residual errors, could prove powerful.

For the automotive industry the particular customer correlation analysis presented here may turn out to be more attractive in the future. As the cars get more and more computerized and intelligent there may be fatigue load-related auxiliary data available for free, at the regular maintenance service stops. A simple download of objective digital data, from each car on the market, would in that case replace or complement the questionnaire survey. For example, GPS data (geographic positioning data) may be used to divide the total mileage into covered distances on various road types.

2 Customer correlation overview

An overview of the customer correlation approach, influenced by people at the fatigue engineering company LMS Durability Technologies, will be presented here. This particular approach was originally suggested for use in the design process of a ground vehicle and will also in our view be considered as such, although applicable to most fatigue-loaded products.

Customer correlation stands for a methodology for specification of the fatigue load environment corresponding to a certain period of time or mileage of severe customer usage, through acquisition and analysis of fatigue load data from the customer environment. The reason for its existence is the desire to base the fatigue design requirements more on what the product is experiencing in real life, rather than on conservative judgements by experience. The latter strategy, on which the automotive industry still more or less relies, leads to needless, costly component modifications and overdesign in general. Usage severity is normally assessed as some extreme population quantile, such as the 95th or 99th percentile.

The characteristic property of our customer correlation approach is the use of a questionnaire survey, for one part of the customer environment measurement. A questionnaire survey makes it possible to inexpensively acquire data, on some more or less fatigue-related auxiliary variables (customer and usage characteristics), from a relatively large customer sample. Assuming that there exists a strong relation between the questionnaire answers and the fatigue load environment, this relation is estimated and used to make the large-sample questionnaire data improve the statistical confidence on the final fatigue load population distribution estimate.

2.1 Customer fatigue load measurements

As mentioned in the introduction, our customer correlation method makes use of two different types of customer fatigue load measurements:

- One direct field measurement of suitable forces, material strains and possibly other usage-revealing entities is made on a small random sample of the customer/product population.
- One less precise, indirect fatigue load measurement is made on a large customer sample (if not the full population), in which the small sample is contained. Fatigue load-related auxiliary data (customer/usage characteristics) are in this case collected, with the help of a questionnaire survey.

The direct measurement is made during a long enough period of time to reduce the random error in the time- (or mileage-)average evaluation of the data. For the case of an automobile, perhaps one year may be an appropriate measurement duration. Many different channels should be measured, in order to capture as much as possible of the complex fatigue-associated loading behavior.

Because of the many signals and the long measurement time, the data storage capability is likely to be insufficient for complete time-series data acquisition. Hence, data reduction, with minimum loss of fatigue-related information, is applied already when collecting data. In fatigue load analysis of force or strain data, the most common means for such a data reduction is extraction and accumulation of rainflow cycles (RFC) into a rainflow cycle matrix (or histogram). Other examples of reduced data formats could be level-crossing diagrams or “time at level” histograms, when the measured signals are not directly related to material strain.

Further data reduction is applied later, in the post-processing of the measured data. For comparison of fatigue load characteristics between different usage situations, etc., a reduction to a manageable set of load variables is necessary, still without truncating the spanning load space too much. One example of such a data reduction is the cycle-by-cycle accumulation of a fatigue damage measure, from the cycles in a RFC matrix. For each cycle i in the matrix, with nominal stress amplitude S_i, a damage value D_i can be calculated by using an S-N curve (or Wöhler curve). The damage measure D results from a summation over all cycles, according to the Palmgren-Miner rule (*Palmgren*, 1924; *Miner*, 1945):

$$D = \sum_i D_i = \sum_i \frac{1}{N_{S_i}}, \tag{2.1}$$

where N_{S_i} is the cycle life, at corresponding constant stress amplitude S_i, obtained from the S-N curve. The fatigue life is considered exhausted when D reaches unity. For comparative fatigue load studies, only the slope of the S-N curve is important (and possibly an endurance limit) and entities proportional to stress, such as strain or force, may be used instead. In our perspective, the fatigue load variables are measures used only to put different fatigue load sequences on a relative fatigue damage impact scale.

Palmgren-Miner damage summation of RFCs is a well-established method, in metal fatigue engineering, to attain a decent estimate of the fatigue damage impact for a given fatigue load history. Also when absolute fatigue life predictions are concerned, it has been shown to yield relatively accurate results (*Dowling*, 1972).

The RFC concept was first proposed in 1967 by Endo (*Matsuishi & Endo*, 1968). Today there exist several more or less equivalent algorithms for RFC counting: See for example *Dowling* & *Socie* (1982); *Rychlik* (1987); *Dreßler et al*. (1997) and *Johannesson* (1999).

To summarize the processing of force and strain data; one measured load signal is instantaneously reduced to a RFC histogram and, in the post-processing, reduced further to a scalar-valued fatigue load variable, whose statistical population distribution (or a specific quantile) is later to be estimated.

The size of the small customer sample, for the direct field measurement, is a trade-off between quality (or inference confidence) and cost. Ideally, the direct measurement should be made on the large customer sample, in which case the questionnaire survey obviously would add no further value to the estimation of load variable distributions.

As with all questionnaire surveys, the choice of questions and their formulations are very important. In order to ensure accurate answers, a survey with personal interviews is desirable. The questionnaire data can be both continuous and categorical. Despite the vital importance of how the questionnaire survey is designed and the processing of questionnaire data, here this issue is left open for later studies and a good survey result is assumed. Another interesting possibility, which should attract attention, is to use some of the auxiliary characteristics for a stratification of the customer population, before the small sample for the direct field measurement is drawn.

2.2 Reference fatigue load measurement

Preceding the customer fatigue load measurement a reference fatigue load measurement is performed, by engineers on a test vehicle. This initial measurement is more comprehensive in terms of data channels and is supposed to cover the intricate fatigue load space more or less completely. In contrast to the customer measurement, the reference measurement is made on many different well-controlled fatigue loading environments separately. For an automobile, examples of such measurement events could be “five runs on test track section A”, “ten drives over a curb at 30 km/h”, etc.

The purpose of the reference measurement is twofold:

- For optimal choice of the subset of measurement channels and the data reduction to fatigue load variables, to be used in the direct customer fatigue load measurement.
- To create a database of signatures or fingerprints (outcomes of the different fatigue load variables), representing all kinds of different usage situations or vehicle maneuvers (events) imaginable.

A limited number of measurement channels or fatigue load variables, for the subset of the direct customer measurement, is necessary for practical reasons. After analysis of the reference measurement data, a suitable subset of data channels is chosen, such that as much coverage of the fatigue load space as possible is retained. Suppose for example that there is a group of suggested load variables with strong correlation, in which case only one of these variables is picked for the subset.

The database of event signatures will later be used when the estimated multivariate fatigue load quantile, representing a certain severe customer fatigue load environment, is converted from numerical values to a more universally interpretable fatigue load specification. This product-independent load specification is composed of a combination of events (with multiplicities) from the database, which makes the result useful for the design of similar products and not only for the particular model used in the investigation (which is already on the market, of course).

2.3 Statistical modelling and inference

After both the direct customer fatigue load measurement and the questionnaire survey are made, the reduced fatigue load variable dependence on the auxiliary variables, measured by the questionnaire, is estimated from the small customer sample data. The dependence is modelled by a multivariate regression model or a general linear model. Alternatively, each dependent variable is modelled separately, using univariate regression models. Normal-distributed residual error terms simplify the analysis in both cases. It is difficult to find the best choice of auxiliary or independent variables to be included in the model and in what form. The auxiliary variables can originally be of either continuous or categorical type, but the analysis may also benefit from conversion of a continuous variable to a categorical one, by discretization. If, for instance, the influence on the dependent fatigue load variables is non-linear and one finds it difficult to make the dependence approximately linear, by transformation of the independent variable, the discretization may be justified.

The model coefficients are estimated in a standard maximum likelihood fashion, repeatedly with different sizes and combinations of independent variables. For high precision in the coefficient estimates, a down-sized model is necessary because of the small sample size. At the same time the model needs to be intricate enough to be able to explain enough variation of the dependent variables. This dilemma is typical for most regression-model selection procedures.

In whatever way it is found, the best model is then used to predict the primary fatigue load measures for customers from which only questionnaire answers are available. Using all the questionnaire data, an empirical, multivariate distribution of the fatigue load variables is calculated.

2.4 Fatigue load specification

The fatigue load distribution estimate, or any quantile values derived from it, is not very useful by itself, as a fatigue load specification. The numerical result has a meaning only for the particular product or vehicle used in the investigation. In order to make the load specification more universal and less product-specific, the numerical result must be associated with some external loading environment. For an automobile, for instance, road profile data of a test track are universal but wheel spindle force data are not. The association to universal load measures is the main objective for the reference fatigue load measurement of loading event signatures.

An optimization software is used to test many different combinations of superposed loading events (with multiplicities) from the event database, while minimizing some calculated measure of distance to the estimated severe target environment. When the best possible match is reached the final fatigue load specification, corresponding to a certain fatigue load severity or population quantile, has the desired form of superposed, external loading events.

3 Problem statement

In order to be able to use questionnaires successfully for the measurement of fatigue load-related, auxiliary information, there has to be a strong relation between how a customer answers the questions and what fatigue load measures that would be expected if a direct measurement (e.g. of force or strain) were to take place. The identification of this relation was briefly discussed in the previous section. Linear regression analysis is used to investigate the requirement needed on such a relation, to justify the method for varying regression-model size, coefficient of determination and size of the small sample. The estimation confidence improvement from the additional auxiliary data is assessed in relation to the case when only the direct load measurement data, from the small customer sample, are used.

Consider the linear regression model,

$$Y = \alpha + \beta' X + \varepsilon, \tag{3.1}$$

where ε is a zero-mean, random error or residual term. Further, assume we have access to n observations (x_i, y_i), from an independently and identically distributed random sample, and additionally N (very large) observations of X only.

In the following, only one scalar fatigue load variable Y, as the dependent (or response) variable, is considered. The method evaluation is performed in two steps. First, a simple regression on one scalar, auxiliary variable, as the independent variable X (or predictor), is studied. Later, the analysis is extended to the case with several questionnaire questions and, thus, a multidimensional auxiliary variable X.

Now, the main issue is when the N additional observations of X will be valuable for inference statements about Y. Typically, we would like to find the qth distribution quantile y_q such that

$$\mathrm{P}(Y \leq y_q) = q, \tag{3.2}$$

for a chosen value of q of, for instance, 0.95 or 0.99.

The possible precision improvement is evaluated in this paper only for the simplified case of normal-distributed auxiliary variables *X*. Even though this in reality would be far from the truth, it is still interesting to study the method performance for this case. Further, a normal-distributed random error term is assumed. Thus, the normal distribution is inherited by the fatigue load variable *Y* itself and, consequently, completely specified by the expectation and variance parameters, μ_Y and σ_Y^2, only. Also the quantile y_q may be explicitly expressed as a function of these parameters:

$$y_q = \mu_Y + z_q \sigma_Y , \tag{3.3}$$

where z_q is the corresponding quantile for the standardized normal distribution $N(0,1)$ and $\sigma_Y = \sqrt{\sigma_Y^2}$ is the standard deviation of *Y*.

Since the normal distribution is imposed on the auxiliary variable *X*, only estimators of the mean μ_X and variance σ_X^2 (or covariance Σ_X, in the case of multivariate *X*) are used to improve inference statements about *Y*. Actually, in the following analyses these distribution parameters are assumed to be known exactly, which represents the idealized, limiting condition when $N \to \infty$. The assumption is motivated by the large sample of additional measurements of *X* and the interest in the method performance when the best possible precision improvement result is achieved.

4 Regression on a scalar, normal auxiliary variable

First, the use for the additional N observations of X is investigated, when X is a scalar random variable. Further, a jointly normal distribution is assumed for X and Y and, hence, Y may be expressed as in Equation 4.1. The random variable (r.v.) U denotes a standardized normal r.v., independent of X, and ρ is the correlation coefficient between X and Y. By comparison of this expression to the regression model in Equation 3.1, the parameters and the random error term of the model are easily identified. See Equation 4.2.

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_X^2 \end{bmatrix} \right), \tag{4.1}$$

$$Y = \mu_Y + \rho\sigma_Y \frac{X - \mu_X}{\sigma_X} + \sigma_Y\sqrt{1-\rho^2}\cdot U\,.$$

$$\begin{aligned} \varepsilon &= \sigma_Y\sqrt{1-\rho^2}\cdot U\,, \\ \beta &= \rho\frac{\sigma_Y}{\sigma_X}\,, \\ \alpha &= \mu_Y - \beta\mu_X\,. \end{aligned} \tag{4.2}$$

Of course, information about X is of no use if X and Y are uncorrelated, i.e. if $\rho = 0$ (and hence $\beta = 0$). In the other extreme situation, when $\rho^2 = 1$, the random error term disappears and the linear regression function is obtained exactly from only two observation pairs. Hence, inferences about X, from the large set of N observations, can easily be transferred to equally well-made inferences about Y. Also for intermediate correlation coefficients, better estimates of μ_X and σ_X, from the large data set, may be useful to improve knowledge about Y. However, as ρ decreases so does the possibility of improvement.

Since the random error term ε of the regression model has zero-mean, the expectation of Y is

$$\mu_Y = \mathrm{E}[Y] = \alpha + \beta\mu_X\,. \tag{4.3}$$

Conditional on X, we have the regression function

$$\mu_{Y|X} = \mathrm{E}[Y|X] = \alpha + \beta X\,. \tag{4.4}$$

Maximum likelihood estimation of the regression parameters, from the sample of n pairs (x_i, y_i), yields

$$\hat{\alpha}' = \hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n \sim N\left(\alpha + \beta\bar{x}_n, \frac{\sigma_\varepsilon^2}{n}\right) \tag{4.5}$$

and

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \sim N\left(\beta, \frac{\sigma_\varepsilon^2}{S_{xx}}\right), \tag{4.6}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \text{ and } S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n). \tag{4.7}$$

4.1 Estimation of response mean

When the true distribution of X is known (or at least very accurately estimated through the large data set of additional observations) μ_Y may be estimated, using the regression model, as

$$\hat{\mu_Y} = \hat{\alpha} + \hat{\beta}\mu_X = \bar{y}_n + \frac{S_{xy}}{S_{xx}}(\mu_X - \bar{x}_n). \tag{4.8}$$

The precision of this estimate is evaluated by the size of its variance, when the small sample is considered unknown. Let $\mathcal{F}$ denote the σ-algebra generated by the r.v. $X_1, X_2, \ldots, X_n$ of the small sample. Then,

$$\mathrm{Var}(\hat{\mu_Y}) = \mathrm{E}[\mathrm{Var}(\hat{\mu_Y} \mid \mathcal{F})] + \mathrm{Var}(\mathrm{E}[\hat{\mu_Y} \mid \mathcal{F}]). \tag{4.9}$$

The second term of this expression is zero, since $\hat{\mu_Y}$ is an unbiased estimate of $\mu_Y = \alpha + \beta\mu_X$ independent of $\mathcal{F}$. Further, since X is assumed to be normal-distributed,

$$\mathrm{Var}(\hat{\mu_Y}) = \mathrm{E}\left[\sigma_\varepsilon^2\left(\frac{1}{n} + \frac{(\mu_X - \bar{X}_n)^2}{S_{XX}}\right)\right] = \frac{\sigma_\varepsilon^2}{n}\left(1 + \frac{\mathrm{E}[F]}{n-1}\right), \tag{4.10}$$

where F denotes an $F(1, n - 1)$-distributed r.v. and the stochastical independence between $\bar{y}_n$ and $\hat{\beta}$, conditional on $\mathcal{F}$, has been used. Note that, with $s_X^2 = S_{XX}/(n-1)$ as the sample variance of $X_1, X_2, \ldots, X_n$,

$$\frac{(\mu_X - \bar{X}_n)^2}{S_{XX}} = \frac{n(\mu_X - \bar{X}_n)^2/\sigma_X^2}{n(n-1)s_X^2/\sigma_X^2} = \frac{1}{n(n-1)}F\,. \qquad \textbf{(4.11)}$$

For $n > 3$, the expectation of F is $(n - 1)/(n - 3)$ and with $\sigma_\varepsilon^2 = \sigma_Y^2(1-\rho^2)$ the variance expression may be rewritten as

$$\mathrm{Var}(\hat{\mu}_Y) = \frac{\sigma_Y^2(1-\rho^2)}{n}\left(1 + \frac{1}{n-3}\right) = \frac{\sigma_Y^2(1-\rho^2)}{n} \cdot \frac{n-2}{n-3}, \quad n > 3\,. \qquad \textbf{(4.12)}$$

Compared with the alternative mean estimate $\bar{y}_n$, as if the regression model and the knowledge about X were not to be used, with variance σ_Y^2/n, the estimation precision is always improved unless both ρ^2 and n are small. More specifically, the precision improves if

$$\rho^2 > \frac{1}{n-2}\,, \qquad n > 3\,. \qquad \textbf{(4.13)}$$

However, for a noticeable improvement when n is not too small, ρ^2 needs to be, say, at least 0.25. Even if the questionnaire survey is less expensive than the direct measurement, the cost must be justified by a sufficient estimation improvement.

4.2 Estimation of response variance

Next, estimates of the variance of Y are considered. If the true distribution of X is known, the regression model can be used to estimate $\sigma_Y^2 = \beta^2\sigma_X^2 + \sigma_\varepsilon^2$ as

$$\hat{\sigma}_Y^2 = \hat{\beta}^2\sigma_X^2 + s_\varepsilon^2\left(1 - \frac{\sigma_X^2}{S_{xx}}\right), \tag{4.14}$$

where the extra term inside the parenthesis makes the estimator unbiased, conditional on $\mathcal{F}$, and

$$s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 \tag{4.15}$$

is an unbiased variance estimator for the random error ε.

Again, the precision of the estimate is evaluated through its variance, when the small sample is considered unknown:

$$\mathrm{Var}(\hat{\sigma}_Y^2) = \mathrm{E}[\mathrm{Var}(\hat{\sigma}_Y^2 \mid \mathcal{F})] + \mathrm{Var}(\mathrm{E}[\hat{\sigma}_Y^2 \mid \mathcal{F}]) \,. \tag{4.16}$$

The estimate $\hat{\sigma}_Y^2$ is unbiased independently of $\mathcal{F}$, since

$$\begin{aligned}\mathrm{E}[\hat{\sigma}_Y^2 \mid \mathcal{F}] &= \sigma_X^2(\mathrm{Var}(\hat{\beta} \mid \mathcal{F}) + \beta^2) + \sigma_\varepsilon^2\left(1 - \frac{\sigma_X^2}{S_{XX}}\right) \\ &= \beta^2\sigma_X^2 + \sigma_\varepsilon^2 = \sigma_Y^2.\end{aligned} \tag{4.17}$$

Hence, the second term in Equation 4.16 is zero. The conditional variance in the first term involves the inverse of a $\chi^2(n-1)$-distributed r.v., denoted below by χ^2, since

$$\begin{aligned}\mathrm{Var}(\hat{\beta}^2 \mid \mathcal{F}) &= \mathrm{E}[\hat{\beta}^4 \mid \mathcal{F}] - (\mathrm{E}[\hat{\beta}^2 \mid \mathcal{F}])^2 \\ &= \beta^4 + 6\beta^2\frac{\sigma_\varepsilon^2}{S_{XX}} + 3\left(\frac{\sigma_\varepsilon^2}{S_{XX}}\right)^2 - \left(\beta^2 + \frac{\sigma_\varepsilon^2}{S_{XX}}\right)^2 \\ &= 4\beta^2\frac{\sigma_\varepsilon^2}{S_{XX}} + 2\left(\frac{\sigma_\varepsilon^2}{S_{XX}}\right)^2\end{aligned} \tag{4.18}$$

and

$$\frac{\sigma_\varepsilon^2}{S_{XX}} = \frac{\sigma_\varepsilon^2}{\sigma_X^2} \cdot \frac{\sigma_X^2}{(n-1)s_X^2} = \frac{\sigma_\varepsilon^2}{\sigma_X^2} \cdot \frac{1}{\chi^2}. \tag{4.19}$$

It is easily verified that a $N(\mu, \sigma^2)$-distributed r.v. has second and fourth moments $\mu^2 + \sigma^2$ and $\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$, respectively. Further, let $2m = n - 1$. Using integration by parts together with $\Gamma(m)$, as the gamma function

$$\Gamma(m) = \int_0^\infty u^{m-1} e^{-u} du \quad , \qquad m > 0 , \tag{4.20}$$

the first two moments of $1/\chi^2$ are calculated as

$$\begin{aligned} \mathrm{E}\left[\frac{1}{\chi^2}\right] &= \int_0^\infty \frac{1}{u} \cdot \frac{u^{m-1}}{2^m \Gamma(m)} e^{-\frac{u}{2}} du \\ &= 0 - \int_0^\infty \frac{u^{m-1}}{(m-1)2^m \Gamma(m)} \left(-\frac{1}{2} e^{-\frac{u}{2}}\right) du \\ &= \frac{1}{2(m-1)} = \frac{1}{n-3} \quad , \qquad n > 3 \end{aligned} \tag{4.21}$$

and

$$\mathrm{E}\left[\frac{1}{\chi^4}\right] = \frac{1}{2(m-2)} \mathrm{E}\left[\frac{1}{\chi^2}\right] = \frac{1}{(n-3)(n-5)} \quad , \qquad n > 5 . \tag{4.22}$$

The variance of the residual variance estimator s_ε^2 is

$$\mathrm{Var}(s_\varepsilon^2 \,|\, \mathcal{F}) = \frac{\sigma_\varepsilon^4}{(n-2)^2} \mathrm{Var}\left(\frac{(n-2)s_\varepsilon^2}{\sigma_\varepsilon^2} \,\middle|\, \mathcal{F} \right) = \frac{2\sigma_\varepsilon^4}{n-2}. \tag{4.23}$$

Hence, since $\hat{\beta}$ and s_ε^2 are stochastically independent, conditional on $\mathcal{F}$, Equation 4.16 expands to

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}_Y^2) &= \mathrm{E}\left[\sigma_X^4 \mathrm{Var}(\hat{\beta}^2 \,|\, \mathcal{F}) + \left(1 - \frac{\sigma_X^2}{S_{XX}}\right)^2 \mathrm{Var}(s_\varepsilon^2 \,|\, \mathcal{F})\right] \\
&= \frac{4\beta^2\sigma_X^2\sigma_\varepsilon^2}{n-3} + \frac{2\sigma_\varepsilon^4}{(n-3)(n-5)} + \frac{2\sigma_\varepsilon^4}{n-2}\left(1 - \frac{2}{n-3} + \frac{1}{(n-3)(n-5)}\right) \\
&= \frac{2\sigma_\varepsilon^4}{n-3}\left(\frac{2\beta^2\sigma_X^2}{\sigma_\varepsilon^2} + \frac{1}{n-5} + \frac{n-3}{n-2} - \frac{2}{n-2} + \frac{1}{(n-2)(n-5)}\right) \\
&= \frac{2\sigma_\varepsilon^4}{n-3}\left(\frac{2\beta^2\sigma_X^2}{\sigma_\varepsilon^2} + 1 - \frac{2(n-7)}{(n-2)(n-5)}\right), \qquad n > 5.
\end{aligned} \tag{4.24}$$

Since $\sigma_\varepsilon^2 = \sigma_Y^2(1-\rho^2)$ and $\beta^2 = \rho^2\sigma_Y^2/\sigma_X^2$, the variance expression may be rewritten as

$$\mathrm{Var}(\hat{\sigma}_Y^2) = \frac{2(1-\rho^2)^2\sigma_Y^4}{n-3}\left(\frac{2\rho^2}{1-\rho^2} + 1 - \frac{2(n-7)}{(n-2)(n-5)}\right), \quad n > 5. \tag{4.25}$$

The natural choice of alternative unbiased estimate of σ_Y^2, when the regression model and the additional X data are not used, is the standard sample variance s_Y^2. Since $(n-1)s_Y^2/\sigma_Y^2$ is a $\chi^2(n\text{-}1)$-distributed r.v.,

$$\mathrm{Var}(s_Y^2) = \mathrm{Var}\left(\frac{\sigma_Y^2}{n-1} \cdot \chi^2\right) = \frac{2\sigma_Y^4}{n-1}. \tag{4.26}$$

Hence, for higher precision of the estimate of σ_Y^2, when using the regression model and the known variance of X, we must have

$$(1-\rho^2)^2\left(\frac{2\rho^2}{1-\rho^2} + 1 - \frac{2(n-7)}{(n-2)(n-5)}\right) < \frac{n-3}{n-1}, \quad n > 5. \tag{4.27}$$

The inequality shows that the precision is always improved, also for this estimate, unless ρ is close to zero at the same time as n is small. The estimation improvement, in terms of variance reduction for some combinations of ρ^2 and n, is presented in Table 4.1.

	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
$\rho^2 = 0.25$	-10	1	3	5	5
$\rho^2 = 0.5$	8	19	21	23	24
$\rho^2 = 0.75$	45	52	53	55	55

Table 4.1 *Reduction of estimator variance in percent, when estimating σ_Y^2 by using the regression model and knowledge about X. A negative value means increase in variance.*

4.3 Estimation of distribution quantile

If a small bias is acceptable, the *q*bo8390

quantile y_q of the fatigue load distribution, from Equation 3.3, may be estimated as

$$\hat{y}_q = \hat{\mu}_Y + z_q\hat{\sigma}_Y = \hat{\mu}_Y + z_q\sqrt{\hat{\sigma}_Y^2}. \tag{4.28}$$

The precision of this biased estimate is evaluated through the mean square error (MSE):

$$\begin{aligned}
\mathrm{MSE}(\hat{y}_q) &= \mathrm{E}[(\hat{y}_q - y_q)^2] \\
&= \mathrm{E}[(\hat{\mu}_Y - \mu_Y)^2 + z_q^2(\hat{\sigma}_Y - \sigma_Y)^2 + 2z_q(\hat{\mu}_Y - \mu_Y)(\hat{\sigma}_Y - \sigma_Y)] \\
&\approx \mathrm{Var}(\hat{\mu}_Y) + \frac{z_q^2}{4\sigma_Y^2}\mathrm{Var}(\hat{\sigma}_Y^2) \\
&= \frac{\sigma_Y^2(1-\rho^2)}{n-3}\left(\frac{n-2}{n} + z_q^2\left(\rho^2 + \frac{1-\rho^2}{2}\left(1 - \frac{2(n-7)}{(n-2)(n-5)}\right)\right)\right).
\end{aligned} \tag{4.29}$$

In the third equality of Equation 4.29 two approximations are involved. First,

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}_Y^2) &= \mathrm{E}[(\hat{\sigma}_Y - \sigma_Y)^2(\hat{\sigma}_Y + \sigma_Y)^2] \\ &= 4\sigma_Y^2\mathrm{E}\left[(\hat{\sigma}_Y - \sigma_Y)^2\left(1 + \frac{1}{2\sigma_Y}(\hat{\sigma}_Y - \sigma_Y)\right)^2\right] \\ &\approx 4\sigma_Y^2\mathrm{E}[(\hat{\sigma}_Y - \sigma_Y)^2] \,.\end{aligned} \tag{4.30}$$

Also, the expectation of the cross-product term of Equation 4.29 is approximately zero, since

$$\begin{aligned}\mathrm{E}[(\hat{\mu}_Y - \mu_Y)(\hat{\sigma}_Y - \sigma_Y)] &\approx \frac{1}{2\sigma_Y}\mathrm{E}[(\hat{\mu}_Y - \mu_Y)(\hat{\sigma}_Y^2 - \sigma_Y^2)] \\ &= \frac{1}{2\sigma_Y}\mathrm{E}[\mathrm{E}[(\hat{\mu}_Y - \mu_Y)(\hat{\sigma}_Y^2 - \sigma_Y^2)\,|\,\mathcal{F}]]\end{aligned} \tag{4.31}$$

and

$$\begin{aligned}&\mathrm{E}[(\hat{\mu}_Y - \mu_Y)(\hat{\sigma}_Y^2 - \sigma_Y^2)\,|\,\mathcal{F}] = \\ &= \mathrm{E}\left[(\bar{Y}_n - \alpha - \beta\bar{X}_n + (\hat{\beta} - \beta)(\mu_X - \bar{X}_n))\left(\hat{\beta}^2\sigma_X^2 + s_\varepsilon^2\left(1 - \frac{\sigma_X^2}{S_{XX}}\right) - \sigma_Y^2\right)\middle|\,\mathcal{F}\right] \\ &= (\mu_X - \bar{X}_n)\sigma_X^2\mathrm{E}[\hat{\beta}^3 - \beta\hat{\beta}^2\,|\,\mathcal{F}] \\ &= (\mu_X - \bar{X}_n)\sigma_X^2\left(\beta^3 + 3\beta\frac{\sigma_\varepsilon^2}{S_{XX}} - \beta\left(\beta^2 + \frac{\sigma_\varepsilon^2}{S_{XX}}\right)\right) \\ &= 2\beta(\mu_X - \bar{X}_n)\frac{\sigma_X^2\sigma_\varepsilon^2}{(n-1)s_X^2}\,,\end{aligned} \tag{4.32}$$

with expectation zero. The second equality in Equation 4.32 is clear since, conditional on $\mathcal{F}$, $\bar{Y}_n$, $\hat{\beta}$ and s_ε^2 are all stochastically independent, unbiased estimators for $\alpha + \beta\bar{X}_n$, β and σ_ε^2, respectively.

Without the additional X data, the quantile y_q would be estimated as

$$\tilde{y}_q = \bar{y}_n + z_q s_Y\,, \tag{4.33}$$

with MSE

$$\begin{aligned}\mathrm{E}[(\tilde{y}_q - y_q)^2] &= \mathrm{E}[(\bar{y}_n - \mu_Y)^2 + z_q^2(s_Y - \sigma_Y)^2] \\ &\approx \mathrm{Var}(\bar{y}_n) + \frac{z_q^2}{4\sigma_Y^2}\mathrm{Var}(s_Y^2) = \sigma_Y^2\left(\frac{1}{n} + \frac{z_q^2}{2(n-1)}\right).\end{aligned} \tag{4.34}$$

The precision improvement of the quantile estimator, when the regression model and the additional *X* data are used, is presented in Table 4.2. It is calculated for the quantiles 0.95 and 0.99 in terms of approximate MSE reduction, for some combinations of ρ^2 and *n*. The result shows that the coefficient of determination ρ^2 has to be almost one half, if a MSE decrease of at least 25% is required. The requirement on ρ^2 becomes stronger for very small sample sizes *n*.

	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
$\rho^2 = 0.25$	0 (-4)	9 (6)	11 (8)	12 (10)	13 (10)
$\rho^2 = 0.5$	22 (17)	30 (26)	32 (28)	34 (30)	35 (31)
$\rho^2 = 0.75$	56 (52)	61 (57)	62 (59)	63 (60)	64 (61)

Table 4.2 *Approximate MSE reduction in percent when estimating the distribution quantile y_q, for q = 0.95 and q = 0.99 (the latter result in parenthesis), by using the regression model and knowledge about X. A negative value means increase in MSE.*

The bias of the quantile estimate $\hat{y}_q$ equals z_q times the bias of $\hat{\sigma}_Y$, defined as $E[\hat{\sigma}_Y] - \sigma_Y$. For the square root of any unbiased variance estimator, such as $\hat{\sigma}_Y$ or *s*, we can derive the relation

$$\begin{aligned} MSE(\hat{\sigma}_Y) &= Var(\hat{\sigma}_Y) + (Bias(\hat{\sigma}_Y))^2 \\ &= E[\hat{\sigma}_Y^2] - (E[\hat{\sigma}_Y])^2 + (E[\hat{\sigma}_Y] - \sigma_Y)^2 \\ &= -2\sigma_Y Bias(\hat{\sigma}_Y) \, . \end{aligned} \quad \textbf{(4.35)}$$

Hence, the bias, as defined above, is negative, with an absolute value proportional to the MSE. Since the MSE of $\hat{\sigma}_Y$ is not easy to calculate, we may instead use the variance of $\hat{\sigma}_Y^2$ from Equation 4.25 and the approximation $MSE(\hat{\sigma}_Y) \approx Var(\hat{\sigma}_Y^2)/(4\sigma_Y^2)$ from Equation 4.30. The resulting bias expression becomes

$$Bias(\hat{\sigma}_Y) \approx -\frac{Var(\hat{\sigma}_Y^2)}{8\sigma_Y^3} \quad \textbf{(4.36)}$$

and, consequently,

$$\mathrm{Bias}(\hat{y}_q) \approx -z_q \frac{\mathrm{Var}(\hat{\sigma}_Y^2)}{8\sigma_Y^3}. \tag{4.37}$$

Hence, provided the approximation is not too crude, use of the knowledge about *X* together with the regression model also reduces the amount of bias, when the estimation precision is improved. In order to substantiate this conclusion, the decisive approximation crudeness is investigated in the following.

The approximation error of Equation 4.30 may be calculated exactly for the standard deviation estimator s_Y. First, the bias of s_Y is derived.

Since, with $2m = n - 1$, $\chi^2 = 2ms^2/\sigma^2$ is $\chi^2(2m)$-distributed, it is also gamma-distributed as $\Gamma(m, 1/2)$ and, hence, $\chi^2/2 \sim \Gamma(m, 1)$. Thus,

$$\mathrm{E}\left[\sqrt{\frac{\chi^2}{2}}\right] = \int_0^\infty \frac{x^{m-\frac{1}{2}} e^{-x}}{\Gamma(m)} dx = \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \tag{4.38}$$

and, since

$$\mathrm{E}[s_Y] = \sqrt{\frac{2\sigma_Y^2}{n-1}}\,\mathrm{E}\left[\sqrt{\frac{\chi^2}{2}}\right], \tag{4.39}$$

the exact bias expression is

$$\mathrm{Bias}(s_Y) = -\sigma_Y\left(1 - \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right). \tag{4.40}$$

The relative approximation error of $\mathrm{Var}(s_Y^2) \approx 4\sigma_Y^2 \mathrm{MSE}(s_Y)$ is expressed analytically below, as a function of the sample size *n*. In Figure 4.1 it is shown graphically that this error goes to zero as *n* increases and is less than one half percent for $n > 27$.

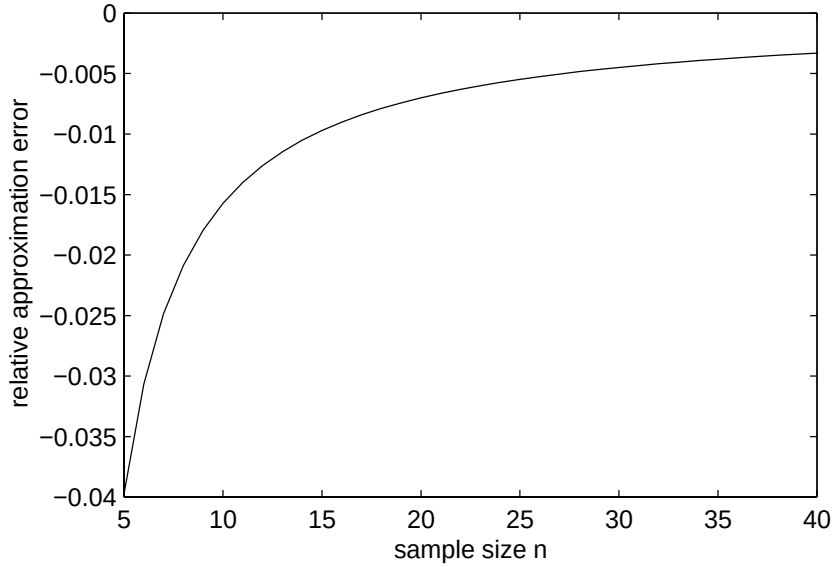

Figure 4.1 *Relative error of the approximation* $\mathrm{Var}(s_Y^2) \approx 4\sigma_Y^2 \mathrm{MSE}(s_Y)$, *defined as in Equation 4.41, as a function of the sample size n.*

$$\text{Relative error} = \frac{4\sigma_Y^2 \mathrm{MSE}(s_Y) - \mathrm{Var}(s_Y^2)}{\mathrm{Var}(s_Y^2)}$$

$$= -\frac{8\sigma_Y^3 \mathrm{Bias}(s_Y)}{\mathrm{Var}(s_Y^2)} - 1 = 4(n-1)\left(1 - \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right) - 1 \,. \tag{4.41}$$

As defined, the relative error appears to be negative, i.e.

$$\frac{\mathrm{Var}(s_Y^2)}{4\sigma_Y^2} > \mathrm{MSE}(s_Y) \,. \tag{4.42}$$

Thus, the approximate expression for the bias of s_Y, equivalent to the expression in Equation 4.36 (for the estimator $\hat{\sigma}_Y$), always overestimates the amount of negative bias. It is reasonable to believe that the relation in Equation 4.42 is also valid with s_Y replaced by $\hat{\sigma}_Y$. Hence, the amount of negative bias of $\hat{\sigma}_Y$ is most likely bounded by the approximate bias of Equation 4.36. Further, a precision improvement for the variance estimate suggests that the approximation error decreases, since the bias is always negative and the bounding variance expression approaches zero. Hence, when the

estimation precision is improved by the regression model and additional X data (i.e. when $\mathrm{Var}(s_Y^2) > \mathrm{Var}(\hat{\sigma}_Y^2)$), it is also reasonable to believe the statement already made that the MSE and therefore also the amount of bias will decrease.

4.4 Alternative regression parameter estimators

Before we go on with the next example and introduce multiple independent variables, it could be argued that our knowledge about X should be used to get a better estimate of β, instead of S_{xy}/S_{xx}, namely

$$\hat{\beta}^* = \frac{S_{xy}}{(n-1)\sigma_X^2}. \tag{4.43}$$

However, conditional on $\mathcal{F}$, this is no longer an unbiased estimate of β. As a consequence, it yields inference statements about Y with lower precision, since the second terms of the variance expressions in both Equation 4.9 and Equation 4.16 are no longer zero. This is shown below only for the alternative mean estimate $\hat{\mu}_Y^* = \bar{Y}_n + \hat{\beta}^*(\mu_X - \bar{X}_n)$.

$$\begin{aligned} \mathrm{E}[\hat{\mu}_Y^* \,|\, \mathcal{F}] &= \mathrm{E}[\bar{Y}_n \,|\, \mathcal{F}] + (\mu_X - \bar{X}_n)\mathrm{E}[\hat{\beta}^* \,|\, \mathcal{F}] \\ &= \alpha + \beta\bar{X}_n + (\mu_X - \bar{X}_n)\frac{\beta s_X^2}{\sigma_X^2} \neq \mu_Y \end{aligned} \tag{4.44}$$

Note that

$$\mathrm{Cov}\left(\bar{X}_n, (\mu_X - \bar{X}_n)\frac{s_X^2}{\sigma_X^2}\right) = -\mathrm{E}[(\bar{X}_n - \mu_X)^2]\mathrm{E}\left[\frac{s_X^2}{\sigma_X^2}\right] = -\mathrm{Var}(\bar{X}_n) \tag{4.45}$$

and

$$\begin{aligned} \mathrm{Var}\left((\mu_X - \bar{X}_n)\frac{s_X^2}{\sigma_X^2}\right) &= \mathrm{E}\left[\left(\frac{\mu_X - \bar{X}_n}{n-1}\right)^2\left(\frac{(n-1)s_X^2}{\sigma_X^2}\right)^2\right] \\ &= \frac{\mathrm{Var}(\bar{X}_n)}{(n-1)^2}(2(n-1) + (n-1)^2) = \frac{n+1}{n-1}\mathrm{Var}(\bar{X}_n), \end{aligned} \tag{4.46}$$

which makes the variance of the expression in Equation 4.44 non-zero for $\rho \neq 0$. Hence,

$$\begin{aligned} \mathrm{Var}(\mathrm{E}[\hat{\mu_Y}^* \,|\, \mathcal{F}]) &= \beta^2 \mathrm{Var}(\bar{X}_n)\left(1 + \frac{n+1}{n-1} - 2\right) \\ &= \frac{2\beta^2\sigma_X^2}{n(n-1)} = \frac{2\rho^2\sigma_Y^2}{n(n-1)}. \end{aligned} \tag{4.47}$$

The first term of the variance of $\hat{\mu_Y}$, from Equation 4.9, does not change much with the different parameter estimator $\hat{\beta}^*$. Since the stochastical independence between $\bar{Y}_n$ and $\hat{\beta}^*$ is still valid, the conditional variance of $\hat{\mu_Y}^*$ is calculated as

$$\begin{aligned} \mathrm{Var}(\hat{\mu_Y}^* \,|\, \mathcal{F}) &= \mathrm{Var}(\bar{Y}_n \,|\, \mathcal{F}) + \left(\frac{\mu_X - \bar{X}_n}{(n-1)\sigma_X^2}\right)^2 \mathrm{Var}(S_{XY} \,|\, \mathcal{F}) \\ &= \frac{\sigma_\varepsilon^2}{n} + \left(\frac{\mu_X - \bar{X}_n}{(n-1)\sigma_X^2}\right)^2 \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \mathrm{Var}(\varepsilon_i \,|\, \mathcal{F}) \\ &= \frac{\sigma_\varepsilon^2}{n}\left(1 + \frac{n(\mu_X - \bar{X}_n)^2}{(n-1)\sigma_X^2} \cdot \frac{s_X^2}{\sigma_X^2}\right), \end{aligned} \tag{4.48}$$

with expectation

$$\mathrm{E}[\mathrm{Var}(\hat{\mu_Y}^* \,|\, \mathcal{F})] = \frac{\sigma_Y^2(1-\rho^2)}{n}\left(1 + \frac{1}{n-1}\right) = \frac{\sigma_Y^2(1-\rho^2)}{n-1}. \tag{4.49}$$

Hence, the total, unconditional variance of $\hat{\mu_Y}^*$ is

$$\begin{aligned} \mathrm{Var}(\hat{\mu_Y}^*) &= \mathrm{E}[\mathrm{Var}(\hat{\mu_Y}^* \,|\, \mathcal{F})] + \mathrm{Var}(\mathrm{E}[\hat{\mu_Y}^* \,|\, \mathcal{F}]) \\ &= \frac{\sigma_Y^2(1-\rho^2)}{n}\left(1 + \frac{1+\rho^2}{(n-1)(1-\rho^2)}\right). \end{aligned} \tag{4.50}$$

Compared to the original result in Equation 4.12, the variance of the alternative mean estimator increases when

$$\frac{1+\rho^2}{(n-1)(1-\rho^2)} > \frac{1}{n-3}, \tag{4.51}$$

i.e. unless both ρ^2 and n are small. The relative increase, defined as

$$\frac{\mathrm{Var}(\hat{\mu}_Y^*) - \mathrm{Var}(\hat{\mu}_Y)}{\mathrm{Var}(\hat{\mu}_Y)} = \frac{n-3}{n-2}\left(1 + \frac{1+\rho^2}{(n-1)(1-\rho^2)}\right) - 1, \qquad \textbf{(4.52)}$$

is shown graphically in Figure 4.2, for $\rho^2 = 0.25$, $\rho^2 = 0.5$ and $\rho^2 = 0.75$.

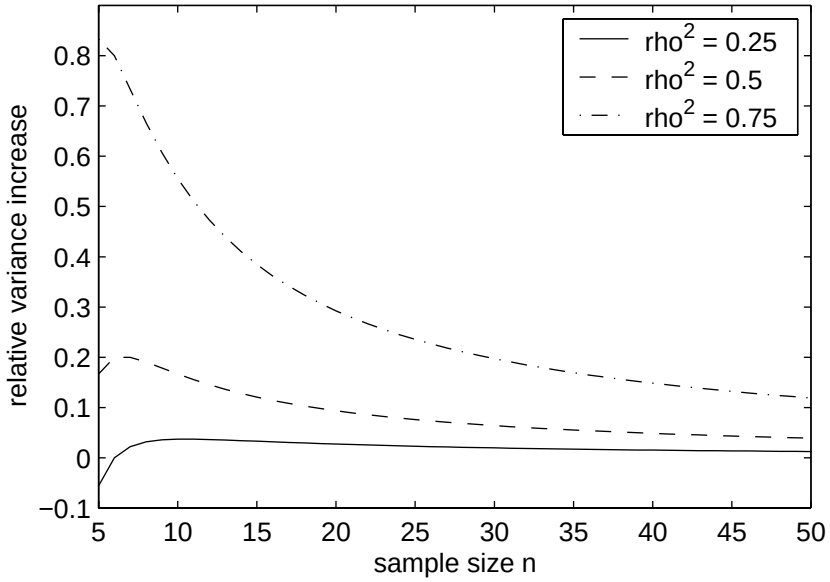

Figure 4.2 *The relative increase in variance, when using the alternative estimator* $\hat{\mu}_Y^*$.

5 Regression on a multivariate, normal auxiliary variable

The previous example is now extended so that several independent variables are considered. The statistical model is from now on a multiple regression model,

$$Y = \alpha + \beta' X + \varepsilon, \tag{5.1}$$

where β and X are column vectors of dimension r and X is multivariate normal-distributed. The error term ε is still a zero-mean, normal-distributed r.v. and, hence,

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \sigma_{XY}' \\ \sigma_{XY} & \Sigma_X \end{bmatrix} \right). \tag{5.2}$$

The $r + 1$ by $r + 1$ covariance matrix is partitioned with respect to the independent and dependent variables, so that Σ_X and σ_{XY} stand for the r by r covariance submatrix of X and the column vector of the covariance between Y and X, respectively. The linear regression function $\mu_{Y|X} = \alpha + \beta' X$, with coefficients

$$\beta = \Sigma_X^{-1} \sigma_{XY} \tag{5.3}$$

and

$$\alpha = \mu_Y - \beta' \mu_X, \tag{5.4}$$

has minimum mean square error $\mathrm{E}[(Y - a - b'X)^2]$, among all linear predictors $a + b'X$ (Johnson & Wichern, 1992). Now, the correlation between Y and this linear regression function is quantified by the multiple correlation coefficient

$$\rho_{Y(X)} = \sqrt{\frac{\sigma_{XY}' \Sigma_X^{-1} \sigma_{XY}}{\sigma_Y^2}}. \tag{5.5}$$

5.1 Estimation of response mean

Since the regression coefficients are not known, they need to be estimated from the small sample of size n. The unbiased maximum likelihood estimator of the regression function is

$$\hat{\mu_{Y|X}} = \bar{y}_n + s_{21}' S_{22}^{-1}(X - \bar{x}_n) = \bar{y}_n + \hat{\beta}'(X - \bar{x}_n), \tag{5.6}$$

where the notation comes from the partition of the unbiased sample estimator for the covariance matrix of $\left[Y \; X'\right]'$, from Equation 5.2,

$$S = \begin{bmatrix} s_Y^2 & s_{21}' \\ s_{21} & S_{22} \end{bmatrix}. \tag{5.7}$$

When the true distribution of X is known, the mean response μ_Y may be estimated as

$$\hat{\mu_Y} = \bar{y}_n + \hat{\beta}'(\mu_X - \bar{x}_n). \tag{5.8}$$

The precision of the μ_Y estimate is again evaluated by its variance. First, note that $\bar{y}_n$ and $\hat{\beta}$ are still stochastically independent and that $\hat{\beta}$ may be written as

$$\begin{aligned} \hat{\beta} &= S_{22}^{-1} s_{21} = S_{22}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta'\bar{x} + \beta'(x_i - \bar{x}) + \varepsilon_i)\right) \\ &= S_{22}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n}((x_i - \bar{x})(x_i - \bar{x})'\beta + (x_i - \bar{x})\varepsilon_i)\right) \\ &= \beta + S_{22}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right). \end{aligned} \tag{5.9}$$

Hence, conditional on $\mathcal{F}$,

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}\,|\,\mathcal{F}) &= \mathrm{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\,|\,\mathcal{F}] \\
&= \frac{1}{(n-1)^2}\mathrm{E}\left[S_{22}^{-1}\left(\sum_{i=1}^{n}(X_i-\bar{X}_n)\varepsilon_i\right)\left(\sum_{j=1}^{n}\varepsilon_j(X_j-\bar{X}_n)'\right)S_{22}^{-1}\,\middle|\,\mathcal{F}\right] \\
&= \frac{1}{(n-1)^2}S_{22}^{-1}\left(\sum_{i=1}^{n}(X_i-\bar{X}_n)\mathrm{E}[\varepsilon_i^2\,|\,\mathcal{F}](X_i-\bar{X}_n)'\right)S_{22}^{-1} \\
&= \frac{\sigma_\varepsilon^2}{n-1}S_{22}^{-1}.
\end{aligned} \tag{5.10}$$

Using Equation 4.9, the variance may now be calculated as

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_Y) &= \mathrm{E}\left[\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{(\mu_X-\bar{X}_n)'S_{22}^{-1}(\mu_X-\bar{X}_n)}{n-1}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{n}\left(1+\frac{\mathrm{E}[T^2]}{n-1}\right),
\end{aligned} \tag{5.11}$$

where T^2 has the Hotelling T^2 distribution with parameters r and n - 1. See *Mardia et al.* (1979) for more on multivariate analysis. Since $F = \{(n-r)/((n-1)r)\}T^2$ has an $F(r, n-r)$ distribution, with mean $(n-r)/(n-r-2)$, and $\sigma_\varepsilon^2 = \sigma_Y^2(1-\rho_{Y(X)}^2)$, the variance expression may be rewritten as

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_Y) &= \frac{\sigma_\varepsilon^2}{n}\left(1+\frac{r\mathrm{E}[F]}{n-r}\right) \\
&= \frac{\sigma_Y^2(1-\rho_{Y(X)}^2)}{n}\cdot\frac{n-2}{n-r-2}, \qquad n>r+2\,.
\end{aligned} \tag{5.12}$$

With $r = 1$ we see that we have the same variance as in Equation 4.12, as expected.

The variance in this example is reduced by the regression model and the knowledge of X, compared with $\mathrm{Var}(\bar{Y}_n) = \sigma_Y^2/n$, if

$$\rho_{Y(X)}^2 > \frac{r}{n-2}\,, \qquad n>r+2\,. \tag{5.13}$$

Hence, for a larger regression model size, a higher multiple correlation coefficient is required.

5.2 Estimation of response variance

Using the multiple regression model of Equation 5.1, the variance of Y is calculated as

$$\sigma_Y^2 = \beta' \Sigma_X \beta + \sigma_\varepsilon^2 \tag{5.14}$$

and one natural choice of variance estimator would be

$$\tilde{\sigma}_Y^2 = \hat{\beta}' \Sigma_X \hat{\beta} + s_\varepsilon^2 . \tag{5.15}$$

However, conditional on $\mathcal{F}$, this estimator is biased. The natural choice of conditional unbiased variance estimator, derived below, is used instead.

One important feature of the trace operator, in linear algebra, is that for two matrices A and B, such that the sizes of A and the transpose of B (or vice versa) are identical,

$$\mathrm{tr}\{AB\} = \mathrm{tr}\{BA\} . \tag{5.16}$$

Using this result, we have

$$\begin{aligned} \mathrm{E}[\hat{\beta}' \Sigma_X \hat{\beta} \,|\, \mathcal{F}] &= \mathrm{tr}\{\Sigma_X \mathrm{E}[\hat{\beta}\hat{\beta}' \,|\, \mathcal{F}]\} \\ &= \mathrm{tr}\{\Sigma_X (\mathrm{Cov}(\hat{\beta} \,|\, \mathcal{F}) + \beta\beta')\} \\ &= \frac{\sigma_\varepsilon^2}{n-1} \mathrm{tr}\{\Sigma_X S_{22}^{-1}\} + \beta' \Sigma_X \beta . \end{aligned} \tag{5.17}$$

Hence, the suggested conditionally unbiased estimator for the variance of Y is

$$\begin{aligned} \hat{\sigma}_Y^2 &= \hat{\beta}' \Sigma_X \hat{\beta} + s_\varepsilon^2 \left(1 - \frac{1}{n-1} \mathrm{tr}\{\Sigma_X S_{22}^{-1}\} \right) \\ &= \hat{\beta}' \Sigma_X \hat{\beta} + s_\varepsilon^2 (1 - \mathrm{tr}\{W^{-1}\}) , \end{aligned} \tag{5.18}$$

where the $r \times r$ random matrix

$$W = \Sigma_X^{-1/2} ((n-1) S_{22}) \Sigma_X^{-1/2} \tag{5.19}$$

has the standardized Wishart distribution $W_r(I_r, n-1)$. Here, as well as in the following, I_r denotes the $r \times r$ identity matrix.

The suggested variance estimator is again evaluated through its variance, as in Equation 4.16. In this multivariate example the calculations require some results on an inverted Wishart-distributed random matrix, for instance the first

two moments. Compare the inverted χ^2-distributed r.v. involved in the simple regression example. The following lemma is given and proved by Das Gupta (1968).

Lemma 5.1 (Das Gupta, 1968): *Let S be an $r \times r$ random matrix distributed according to the Wishart distribution $W_r(\Sigma, m)$, where* $\text{rank}(\Sigma) = r$. *Then,*

$$\begin{aligned} &(a) \quad \mathrm{E}[S] = m\Sigma \\ &(b) \quad \mathrm{E}[S^{-1}] = \frac{1}{m-r-1}\Sigma^{-1}, \text{ if } m > r+1 \\ &(c) \quad \mathrm{E}[S^{-1}\Sigma S^{-1}] = \frac{m-1}{(m-r)(m-r-1)(m-r-3)}\Sigma^{-1}, \ m > r+3 \end{aligned} \tag{5.20}$$

In our calculations the inverted standard Wishart-distributed matrix also appears in the conditional covariance of the random vector $\Sigma_X^{1/2}\hat{\beta}$,

$$\mathrm{Cov}(\Sigma_X^{1/2}\hat{\beta}\,|\,\mathcal{F}) = \frac{\sigma_\varepsilon^2}{n-1}\Sigma_X^{1/2}S_{22}^{-1}\Sigma_X^{1/2} = \sigma_\varepsilon^2 W^{-1}. \tag{5.21}$$

By the spectral decomposition theorem we can write

$$\mathrm{Cov}(\Sigma_X^{1/2}\hat{\beta}\,|\,\mathcal{F}) = \sigma_\varepsilon^2 W^{-1} = \sigma_\varepsilon^2 \Gamma\Lambda^{-1}\Gamma', \tag{5.22}$$

where Λ is a diagonal matrix with positive random elements λ_k and Γ is an orthogonal random matrix. Conditional on $\mathcal{F}$, however, both Λ and Γ are given, as S_{22} is given. With the transformation to a vector with stochastically independent components $\eta = \Gamma'\Sigma_X^{1/2}\hat{\beta}$, conditional on $\mathcal{F}$,

$$\mathrm{Var}(\hat{\beta}'\Sigma_X\hat{\beta}\,|\,\mathcal{F}) = \mathrm{Var}(\eta'\eta\,|\,\mathcal{F}) = \sum_{k=1}^{r}\mathrm{Var}(\eta_k^2\,|\,\mathcal{F}). \tag{5.23}$$

Again, as in Equation 4.18, we have that the variance of a squared $N(\mu, \sigma^2)$-distributed r.v. is $4\mu^2\sigma^2 + 2\sigma^4$. Conditional on $\mathcal{F}$, $\eta \sim N(\mu_\eta, \sigma_\varepsilon^2\Lambda^{-1})$, where $\mu_\eta = \Gamma'\Sigma_X^{1/2}\beta$. Thus,

$$\begin{aligned} \mathrm{Var}(\hat{\beta}'\Sigma_X\hat{\beta}\,|\,\mathcal{F}) &= \sum_{k=1}^{r}(4\sigma_\varepsilon^2\mu_{\eta,k}^2\lambda_k^{-1} + 2\sigma_\varepsilon^4\lambda_k^{-2}) \\ &= 4\sigma_\varepsilon^2\mathrm{tr}\{\mu_\eta\mu_\eta'\Lambda^{-1}\} + 2\sigma_\varepsilon^4\mathrm{tr}\{\Lambda^{-2}\} \\ &= 4\sigma_\varepsilon^2\beta'\Sigma_X^{1/2}W^{-1}\Sigma_X^{1/2}\beta + 2\sigma_\varepsilon^4\mathrm{tr}\{W^{-2}\} \end{aligned} \tag{5.24}$$

with unconditional expectation, if $n > r + 4$,

$$\begin{aligned} \mathrm{E}[\mathrm{Var}(\hat{\beta}'\Sigma_X\hat{\beta}\,|\,\mathcal{F})] &= \frac{2\sigma_\varepsilon^2}{n-r-2}\left(2\beta'\Sigma_X\beta + \frac{(n-2)r\sigma_\varepsilon^2}{(n-r-1)(n-r-4)}\right) \\ &= \frac{2(1-\rho_{Y(X)}^2)^2\sigma_Y^4}{n-r-2}\left(\frac{2\rho_{Y(X)}^2}{1-\rho_{Y(X)}^2} + \frac{(n-2)r}{(n-r-1)(n-r-4)}\right). \end{aligned} \tag{5.25}$$

In Equation 5.25 we have used the equalities $\beta'\Sigma_X\beta = \rho_{Y(X)}^2\sigma_Y^2$ and $\sigma_\varepsilon^2 = (1-\rho_{Y(X)}^2)\sigma_Y^2$.

The following Lemma 5.2 is given and proved by von Rosen (1988) and will be shown useful in the subsequent calculations, involving the second term of the variance estimator in Equation 5.18. However, first the stacking operator (vec), the Kronecker product and the commutation matrix (also called the permuted identity matrix) are defined.

Let $A = [a_{ij}]$ be an $m \times n$ matrix and a_j the jth column of A. Then, vec(A) is the resulting $mn \times 1$ column vector when stacking the columns of A on top of each other;

$$\mathrm{vec}(A) = \left[a_1' \; a_2' \; \ldots \; a_n'\right]'.$$

Further, let B be a $p \times q$ matrix. Then the Kronecker product $A \otimes B$ is defined as the $mp \times nq$ matrix

$$A \otimes B = \left[a_{ij}B\right].$$

The $mn \times mn$ commutation matrix K_{mn} is defined as

$$K_{mn} = \sum_{i=1}^{m}\sum_{j=1}^{n} H_{ij} \otimes H_{ij}',$$

where H_{ij} is an $m \times n$ matrix with a 1 in position ij and zeros elsewhere. Magnus & Neudecker (1979) give useful results on the commutation matrix. For instance,

$$\mathrm{tr}\{K_{mn}\} = 1 + d(m-1, n-1), \tag{5.26}$$

where $d(m, n)$ is the greatest common divisor of m and n. Now we can state the lemma.

Lemma 5.2 (von Rosen, 1988): *Let* $S \sim W_r(\Sigma, m)$. *Then*

$$\begin{aligned} \mathrm{E}[S^{-1} \otimes S^{-1}] &= c_1(\Sigma^{-1} \otimes \Sigma^{-1}) + c_2 \mathrm{vec}(\Sigma^{-1})\mathrm{vec}(\Sigma^{-1})' \\ &\quad + c_2 K_{rr}(\Sigma^{-1} \otimes \Sigma^{-1}) , \end{aligned} \tag{5.27}$$

where $c_2^{-1} = (m-r)(m-r-1)(m-r-3)$ *and* $c_1 = (m-r-2)c_2$.

Using the stochastical independence between $\hat{\beta}$ and s_ε^2, conditional on $\mathcal{F}$, the precision of the variance estimate $\hat{\sigma}_Y^2$ of Equation 5.18 can finally be calculated as

$$\begin{aligned} \mathrm{Var}(\hat{\sigma}_Y^2) &= \mathrm{E}[\mathrm{Var}(\hat{\sigma}_Y^2 \,|\, \mathcal{F})] + \mathrm{Var}(\mathrm{E}[\hat{\sigma}_Y^2 \,|\, \mathcal{F}]) \\ &= \mathrm{E}\left[\mathrm{Var}(\hat{\beta}'\Sigma_X\hat{\beta} \,|\, \mathcal{F}) + \frac{2\sigma_\varepsilon^4}{n-r-1}(1 - \mathrm{tr}\{W^{-1}\})^2 \right] . \end{aligned} \tag{5.28}$$

In our case $W \sim W_r(I_r, n-1)$ and for $n > r+4$ we have, since $(\mathrm{tr}\{W^{-1}\})^2 = \mathrm{tr}\{W^{-1} \otimes W^{-1}\}$,

$$\begin{aligned} \mathrm{E}[(1 - \mathrm{tr}\{W^{-1}\})^2] &= 1 - 2\mathrm{tr}\{\mathrm{E}[W^{-1}]\} + \mathrm{tr}\{\mathrm{E}[W^{-1} \otimes W^{-1}]\} \\ &= 1 - \frac{2r}{n-r-2} + \mathrm{tr}\{c_1 I_{r^2} + c_2 \mathrm{vec}(I_r)\mathrm{vec}(I_r)' + c_2 K_{rr} I_{r^2}\} \\ &= 1 - \frac{2r}{n-r-2} + c_1 r^2 + c_2 r + c_2 r \\ &= 1 - \frac{2r}{n-r-2} + \frac{(n-r-3)r^2 + 2r}{(n-r-1)(n-r-2)(n-r-4)} . \end{aligned} \tag{5.29}$$

Hence, for $n > r+4$,

$$\begin{aligned} \mathrm{Var}(\hat{\sigma}_Y^2) &= \frac{2(1-\rho_{Y(X)}^2)^2\sigma_Y^4}{n-r-2}\left(\frac{2\rho_{Y(X)}^2}{1-\rho_{Y(X)}^2} + \frac{(n-2)r}{(n-r-1)(n-r-4)} \right) \\ &\quad + \frac{2(1-\rho_{Y(X)}^2)^2\sigma_Y^4}{n-r-2}\left(\frac{n-3r-2}{n-r-1} + \frac{(n-r-3)r^2 + 2r}{(n-r-1)^2(n-r-4)} \right) \\ &= \frac{2(1-\rho_{Y(X)}^2)^2\sigma_Y^4}{n-r-2}\left(\frac{2\rho_{Y(X)}^2}{1-\rho_{Y(X)}^2} + 1 - a(n,r) \right) , \end{aligned} \tag{5.30}$$

where

$$a(n, r) = \frac{(r+1)(n^2-(4r+5)n+3r(r+3))+4}{(n-r-1)^2(n-r-4)} . \quad \textbf{(5.31)}$$

Notice that for $r = 1$ (and hence $\rho^2_{Y(X)} = \rho^2$) the precision expression reduces to the simple regression expression in Equation 4.25, as expected, since

$$a(n, 1) = \frac{2(n-7)}{(n-2)(n-5)} . \quad \textbf{(5.32)}$$

For higher precision in the proposed estimator for σ^2_Y (again compared to s^2_Y), when using the multiple regression model and the known covariance of X, we must have

$$(1-\rho^2_{Y(X)})^2\left(\frac{2\rho^2_{Y(X)}}{1-\rho^2_{Y(X)}}+1-a(n, r)\right) < \frac{n-r-2}{n-1}, \quad n > r+4 . \quad \textbf{(5.33)}$$

The improvements in terms of variance reductions as a function of model size r, compared to $\mathrm{Var}(s^2_Y)$ when the X data are not used, are shown graphically in Figure 5.1, for $\rho^2_{Y(X)} = 0.5$, $\rho^2_{Y(X)} = 0.75$ and for two different sample sizes, $n = 30$ and $n = 100$. For an optimal variance reduction and for fixed n, it is shown that there has to be a trade-off between a high multiple correlation coefficient and a small regression model size, since the correlation increases with the model size.

Note that the model size r here refers to the dimension of X or, equivalently, the number of predictors. A different and maybe more common interpretation of regression model size is the number of regression parameters, which in our case is $r + 1$ with the parameter α added.

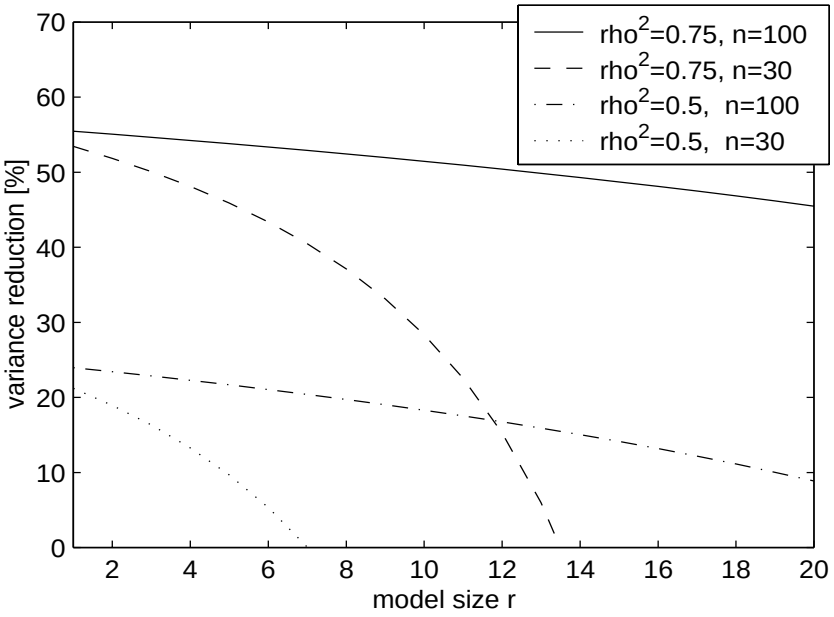

Figure 5.1 *Variance reduction, in percent, for the proposed variance estimator $\hat{\sigma}_Y^2$, as a function of the regression model size, compared to when only Y data are analyzed.*

5.3 Estimation of distribution quantile

Finally, the quantile y_q is estimated as before in Equation 4.28, but with the new mean and variance estimators. The precision of the quantile estimate is evaluated through the approximate MSE. The same approximations as in Equation 4.29 are still valid. For instance, the expectation of the cross-product term vanishes approximately also in this multivariate case. Compare Equation 4.31 and Equation 4.32.

$$
\begin{aligned}
&\mathrm{E}[(\hat{\mu}_Y - \mu_Y)(\hat{\sigma}_Y^2 - \sigma_Y^2)\,|\,\mathcal{F}] \;= \\
&= \mathrm{E}[(\hat{\beta}-\beta)'(\mu_X - \bar{X}_n)\cdot(\hat{\beta}'\Sigma_X\hat{\beta} + s_\varepsilon^2(1-\mathrm{tr}\{W^{-1}\}) - \sigma_Y^2)\,|\,\mathcal{F}] \\
&= \mathrm{E}[(\hat{\beta}-\beta)'(\mu_X - \bar{X}_n)\cdot\hat{\beta}'\Sigma_X\hat{\beta}\,|\,\mathcal{F}] \qquad (5.34) \\
&= \mathrm{E}[(\hat{\beta}-\beta)'(\mu_X - \bar{X}_n)\cdot(\hat{\beta}-\beta)'\Sigma_X(\hat{\beta}-\beta)\,|\,\mathcal{F}] \\
&\quad + 2\mathrm{E}[(\hat{\beta}-\beta)'(\mu_X - \bar{X}_n)\cdot(\hat{\beta}-\beta)'\Sigma_X\beta\,|\,\mathcal{F}]
\end{aligned}
$$

First, the initial term itself, in the resulting expression above, is shown to be zero. With the Mahalanobis transformation of the centered estimator vector $\hat{\beta} - \beta$, or actually a transformation by $S_{22}^{1/2}$ only, the transformed zero-mean vector $z = S_{22}^{1/2}(\hat{\beta} - \beta)$ has, conditional on $\mathcal{F}$, stochastically independent and normal-distributed components. Thus,

$$\begin{aligned} & \mathrm{E}[(\hat{\beta} - \beta)'(\mu_X - \bar{X}_n) \cdot (\hat{\beta} - \beta)'\Sigma_X(\hat{\beta} - \beta) \,|\, \mathcal{F}] \\ &= \mathrm{E}[z' S_{22}^{-1/2}(\mu_X - \bar{X}_n) \cdot z' S_{22}^{-1/2}\Sigma_X S_{22}^{-1/2} z \,|\, \mathcal{F}] \\ &= \mathrm{E}[z'a \cdot z'Bz \,|\, \mathcal{F}] = \sum_{i,\,j,\,k} a_i b_{jk} \mathrm{E}[z_i z_j z_k \,|\, \mathcal{F}] = 0 , \end{aligned} \tag{5.35}$$

where $a = [a_i] = S_{22}^{-1/2}(\mu_X - \bar{X}_n)$ and $B = [b_{jk}] = S_{22}^{-1/2}\Sigma_X S_{22}^{-1/2}$.

The second term in Equation 5.34 has the unconditional expectation zero, since

$$\begin{aligned} & \mathrm{E}[\mathrm{E}[(\hat{\beta} - \beta)'(\mu_X - \bar{X}_n) \cdot (\hat{\beta} - \beta)'\Sigma_X\beta \,|\, \mathcal{F}]] \\ &= \mathrm{E}[(\mu_X - \bar{X}_n)'\mathrm{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \,|\, \mathcal{F}]\Sigma_X\beta] \\ &= \mathrm{E}\left[\frac{\sigma_\varepsilon^2}{n-1}(\mu_X - \bar{X}_n)' S_{22}^{-1}\Sigma_X\beta\right] = 0 . \end{aligned} \tag{5.36}$$

Hence, for $n > r + 4$ and $a(n, r)$ as given in Equation 5.31,

$$\begin{aligned} \mathrm{MSE}(\hat{y}_q) &\approx \mathrm{Var}(\hat{\mu}_Y) + \frac{z_q^2}{4\sigma_Y^2}\mathrm{Var}(\hat{\sigma}_Y^2) \\ &\approx \frac{\sigma_Y^2(1 - \rho_{Y(X)}^2)}{n - r - 2}\left(\frac{(n-2)}{n} + z_q^2\left(\rho_{Y(X)}^2 + \frac{1 - \rho_{Y(X)}^2}{2}(1 - a(n, r))\right)\right) . \end{aligned} \tag{5.37}$$

Compared to the MSE in Equation 4.34, when the additional *X* data are not used, the approximate accuracy improvement for the quantile estimator $\hat{y}_q$ is shown graphically as a function of the model size *r*. The 0.99 quantile function is calculated for two different sample sizes, $n = 30$ and $n = 100$, and two different multiple correlation coefficients, $\rho_{Y(X)}^2 = 0.5$ and $\rho_{Y(X)}^2 = 0.75$. See Figure 5.2. An analogous graphical result for the 0.95 quantile is presented in Figure 5.3. When the method performs well, the less extreme quantile result shows a slightly higher accuracy improvement for similar conditions. This accuracy difference is visualized more clearly in Figure 5.4, for the intermediate sample size $n = 50$.

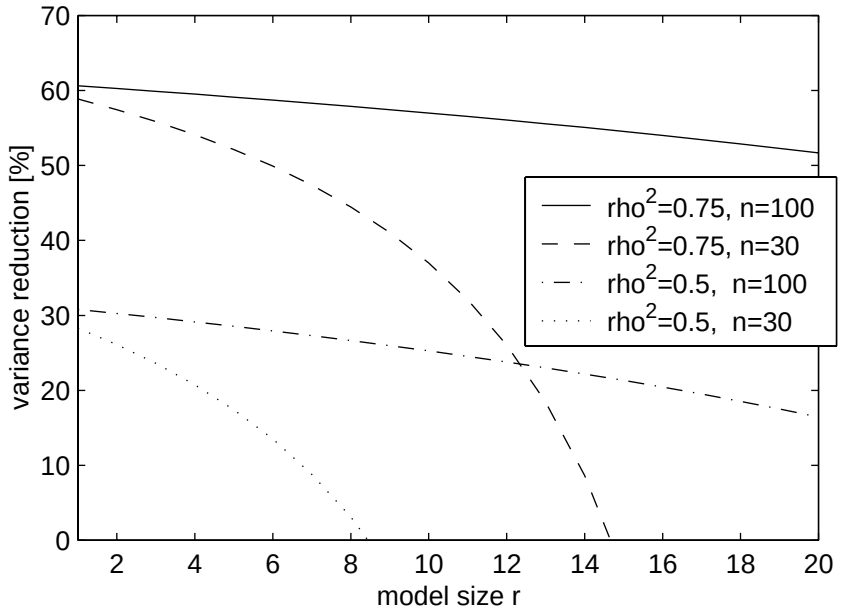

Figure 5.2 *MSE reduction, in percent, for the proposed 0.99 quantile estimator, as a function of the regression model size, compared to when only Y data are analyzed.*

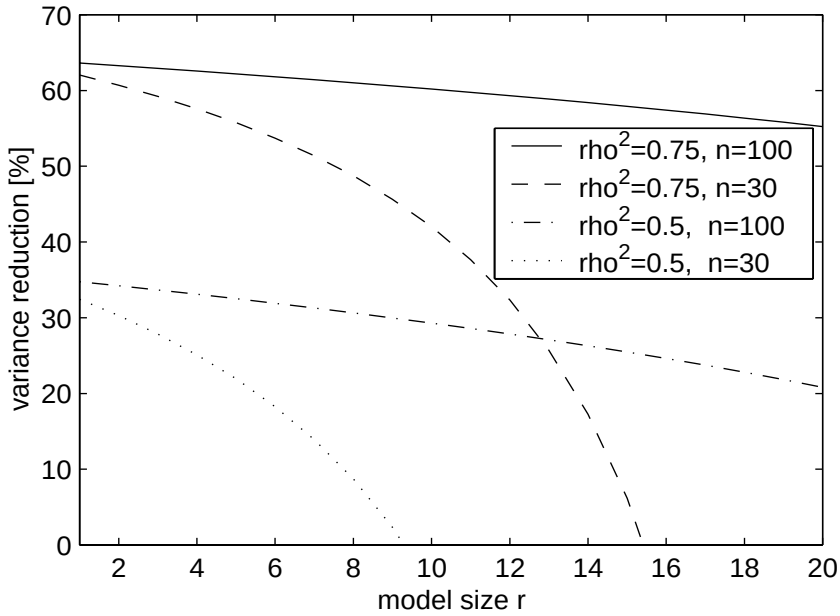

Figure 5.3 *MSE reduction, in percent, for the proposed 0.95 quantile estimator, as a function of the regression model size, compared to when only Y data are analyzed.*

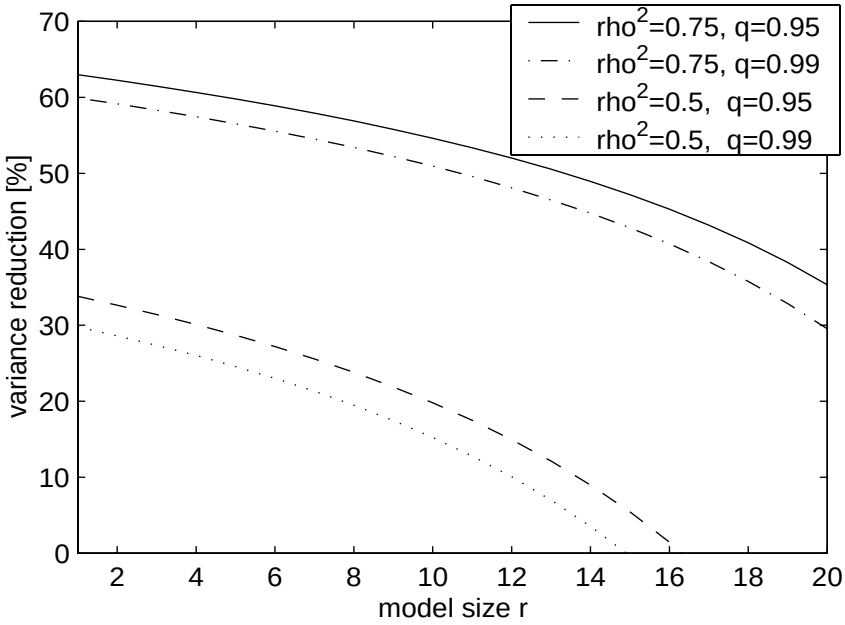

Figure 5.4 *MSE reduction, in percent, for the proposed quantile estimators and sample size n = 50, as a function of the regression model size, compared to when only Y data are analyzed.*

If again a 25% MSE decrease of the quantile estimate is required, to justify the cost of a questionnaire survey, the coefficient of determination $\rho^2_{Y(X)}$ has to be at least one half. However, if the sample size *n* is very small (say $n = 50$ or less), this coefficient of determination must be attained with only a few independent variables in the model.

6 Preliminaries on general auxiliary variable distribution

One natural way to proceed, after the present analysis, would be to release the normal-distribution restriction on the auxiliary variable X. When this restriction is dropped the existing response mean and variance estimators, as well as all $\mathcal{F}$-measurable results, are still valid. However, all the unconditional performance results change and a new quantile estimator must be proposed. For a general statistical distribution $F_N(x)$ of the auxiliary variable X, the qth quantile y_q may be written as

$$y_q = \arg\{y : \mathrm{E}[\mathrm{P}(Y \le y|X)] = q\}. \tag{6.1}$$

Conditional on X, the fatigue load Y is still normal-distributed, as long as the regression model residual ε is normal-distributed. Hence,

$$y_q = \arg\left\{y : \mathrm{E}\left[\Phi\left(\frac{y-\mu_{Y|X}}{\sigma_\varepsilon}\right)\right] = q\right\}, \tag{6.2}$$

where $\mu_{Y|X} = \alpha + \beta X$ is our regression function from Equation 4.4.

Since the true regression function and residual variance is not known, the parameters must be replaced by their point estimates, $\hat{\mu}_{Y|X} = \hat{\alpha} + \hat{\beta}X$ and s_ε, from the small customer sample. However, conditional on X, this replacement makes the normal distribution function $\Phi(u)$ invalid, as

$$\frac{Y-\hat{\mu}_{Y|X}}{s_\varepsilon}\cdot c(X) = \frac{(Y-\mu_{Y|X}+\alpha-\hat{\alpha}+(\beta-\hat{\beta})X)/\sigma_\varepsilon}{s_\varepsilon/\sigma_\varepsilon}\cdot c(X) \tag{6.3}$$

has the student's t distribution with $n-r-1$ degrees of freedom. The function $c(X)$ compensates for the excessive variance of the numerator, as it should be unity to make the expression t-distributed. Since Y is stochastically independent of the small sample data, and therefore also independent of $\hat{\alpha}$ and $\hat{\beta}$,

$$\mathrm{Var}((Y-\hat{\mu}_{Y|X})\big|X) = \sigma_\varepsilon^2\left(1+\frac{1}{n}+\frac{(X-\bar{x}_n)S_{22}^{-1}(X-\bar{x}_n)}{n-1}\right). \tag{6.4}$$

Compare the result in Equation 5.10 and Equation 5.11. Apparently, the function $c(X)$ must equal

$$c(X) = \left(1 + \frac{1}{n} + \frac{(X - \bar{x}_n)S_{22}^{-1}(X - \bar{x}_n)}{n-1}\right)^{-1/2}. \quad (6.5)$$

Consequently, one natural suggestion for estimating the quantile y_q is

$$\hat{y_q} = \arg\left\{y : \int F_t\left(\frac{y - \hat{\mu}_{Y|x}}{s_\varepsilon} \cdot c(x)\right) dF_N(x) = q\right\}, \quad (6.6)$$

where $F_t(u)$ is the cumulative distribution function for the $t(n-r-1)$ distribution.

Taking also the direct load measurement on the small sample $\{y_j\}$ into account, a similar quantile estimator may be formulated as

$$\hat{y_q}^* = \arg\left\{y : \frac{1}{wn+N}\left(N\int F_t\left(\frac{y - \hat{\mu}_{Y|x}}{s_\varepsilon} \cdot c(x)\right) dF_N(x) + w\sum_{j=1}^{n} I_{\{y_j < y\}}\right) = q\right\}, \quad (6.7)$$

where w is a weight constant, N is the size of the full population (or number of questionnaire replies) when the small sample is excluded, and the indicator function

$$I_{\{y_j < y\}} = \begin{cases} 1 & \text{if } y_j < y \\ 0 & \text{if } y_j \geq y \end{cases}. \quad (6.8)$$

When $F_N(x)$ is the non-parametric, empirical distribution of X, Equation 6.7 may be rewritten as

$$\hat{y_q}^* = \arg\left\{y : \frac{1}{wn+N}\left(\sum_{i=1}^{N} F_t\left(\frac{y - \hat{\mu}_{Y|x_i}}{s_\varepsilon} \cdot c(x_i)\right) + w\sum_{j=1}^{n} I_{\{y_j < y\}}\right) = q\right\}. \quad (6.9)$$

If one finds it appealing to add weight to each direct measurement, for accuracy reasons, the weight constant w is assigned a value greater than one. With $w = 1$ all measurements have the same weight.

The precision of the quantile estimators above presumably depend on the auxiliary variable distribution $F_N(x)$. Its evaluation remains as a challenge for the future.

7 Summary and conclusions

Results of estimation precision for population distribution parameters and, more importantly, extreme distribution quantiles have been presented for a univariate r.v. representing a certain scalar fatigue load measure. This fatigue load variable Y is for cost reasons only possible to measure directly on a small sample of the population. In order to improve inference confidence, a less expensive measurement of fatigue load-related, auxiliary variables is performed on a much larger sample. In fact, data on the auxiliary variables are assumed to be extensively acquired on the full population, by means of a questionnaire survey. Hence, the true population distribution of the auxiliary r.v. is assumed known, so our results reflect the estimation precision when the best possible measurement of the auxiliary variables is achieved. The difficult issues regarding the questionnaire design and other questionnaire survey policies are not dealt with here.

As an application we may think of some degree of fatigue loading of an automobile and its distribution over the customer population. Knowledge about the population distribution of such a load variable would be of great value in the design process of a new car model.

A regression analysis is performed to estimate the indispensable influence of the auxiliary data on the fatigue load response variable, using data from the small sample. The random residual error of the regression model is assumed to be normal-distributed. If the relation between the fatigue load variable Y and the vector-valued auxiliary variable X is shown to be strong enough, the estimated model and the auxiliary data are used to include the full population in the inference about the distribution of the fatigue load variable Y. The resulting estimator precision improvement, compared to when only Y data from the small sample are analyzed, is what we use to quantify the method performance. For varying correlation between X and Y and for varying sizes of the small sample and the regression model (dimension of X), the precision improvement is presented graphically. In this report, the result is limited to the case when the auxiliary r.v. X is normal-distributed.

The investigation shows that better estimates about the distribution of Y are possible, provided the questionnaire answers capture the customer fatigue loading behavior well and enough variation in the fatigue load variable can be explained with a multiple regression model of limited model size. In particular when the sample size n is very small, too many independent variables in the regression model lead to insufficient improvement or even a loss of pre-

cision. If a 25% decrease of the mean square error (MSE) of the quantile estimate is required, to justify the cost of a questionnaire survey, the coefficient of determination $\rho^2_{Y(X)}$ has to be at least one half. However, if the sample size is as small as $n = 50$ or less, this coefficient of determination must be attained with only a few independent variables in the model. See Figure 5.2 - Figure 5.4.

One natural way to proceed, after the present analysis, would be to release the normal-distribution restriction on the auxiliary variable X. When this restriction is dropped the existing response mean and variance estimators, as well as all $\mathcal{F}$-measurable results, are still valid. However, the quantile estimator and all unconditional results must be replaced. As before, $\mathcal{F}$ denotes the σ-algebra generated by the auxiliary r.v. in the small sample.

As already mentioned in the introduction, a stratification of the population could be beneficial. A better regression model estimation may be possible if stratified sampling is used, for the direct customer measurement. Further, the extension to several fatigue load variables is natural. One difficulty that most likely would turn up in the arising multivariate regression analysis is the modelling of covariance structure for the random residual vector.

References

Das Gupta, S. 1968. "Some aspects on discrimination function coefficients." *Sankhyã, the Indian Journal of Statistics, Series A.* 30:387-400.

Dowling, N. E. 1972. "Fatigue predictions for complicated stress-strain histories." *Journal of Materials* 7:71-87.

Dowling, N. E. & Socie, D.F. 1982. "Simple rainflow counting algorithms." *International Journal of Fatigue* 4:31-40.

Dreßler, K.; Hack, M. & Krüger, W. 1997. "Stochastic reconstruction of loading histories from a rainflow matrix." *Zeitschrift für Angewandete Mathematik und Mechanik* 77:217-226.

Johnson, R. A. & Wichern, D.W. 1992. *Applied Multivariate Statistical Analysis*, third edition. Englewood Cliffs, USA: Prentice-Hall.

Johannesson, P. 1999. *Rainflow Analysis of Switching Markov Loads*. Doctoral Theses in Mathematical Sciences 1999:4. Lund, Sweden: Lund Institute of Technology.

Magnus, J. R. & Neudecker, H. 1979. "The commutation matrix: Some properties and applications." *The Annals of Statistics* 7:381-394.

Mardia, K. V.; Kent, J. T. & Bibby, J.M. 1979. *Multivariate Analysis*. London: Academic Press.

Matsuishi, M. & Endo, T. 1968. "Fatigue of metals subjected to varying stress." Paper presented to Japan Society of Mechanical Engineers, Jukvoka, Japan.

Miner, M. A. 1945. "Cumulative damage in fatigue." *Journal of Applied Mechanics* 12:A159-A164.

Palmgren, A. 1924. "Die lebensdauer von kugellagern." *Zeitschrift des Vereins Deutscher Ingenieure* 68:339-341.

von Rosen, D. 1988. "Moments for the inverted Wishart distribution." *Scandinavian Journal of Statistics* 15:97-109.

Rychlik, I. 1987. "A new definition of the rainflow cycle counting method." *International Journal of Fatigue* 9:119-121.

Thomas, J. J.; Perroud G.; Bignonnet, A. & Monnet D. 1999. "Fatigue design and reliability in the automotive industry." In Marquis, G. & Solin, J. (editors), *Fatigue Design and Reliability, ESIS Publication 23*. Oxford, UK: Elsevier Science, 1-12.

Wang, J. Z.; Muddiman, M. W. & Moore, G. R. 1999. "Structural correlation of automotive proving grounds to China customer field usage." In Wu, X. R. & Wang, Z.G. (editors) *Fatigue '99, Proceedings of the Seventh International Fatigue Congress, Beijing, P. R. China*. Cradley Heath, UK: EMAS, 2379-2384.