

Kansas State University Libraries

## New Prairie Press

---

Conference on Applied Statistics in Agriculture      2016 - 28th Annual Conference Proceedings

---

# DEVELOPING PREDICTION EQUATIONS FOR FAT FREE LEAN IN THE PRESENCE OF AN UNKNOWN AMOUNT OF PROPORTIONAL MEASUREMENT ERROR

Zachary J. Hass

*Purdue University*, [zhass@purdue.edu](mailto:zhass@purdue.edu)


Bruce A. Craig

*Purdue University*, [bacraig@purdue.edu](mailto:bacraig@purdue.edu)

Allan Schinckel

*Purdue University*, [aschinck@purdue.edu](mailto:aschinck@purdue.edu)

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

 Part of the [Agriculture Commons](#), [Animal Studies Commons](#), [Applied Statistics Commons](#), and the [Meat Science Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

---

### Recommended Citation

Hass, Zachary J.; Craig, Bruce A.; and Schinckel, Allan (2016). "DEVELOPING PREDICTION EQUATIONS FOR FAT FREE LEAN IN THE PRESENCE OF AN UNKNOWN AMOUNT OF PROPORTIONAL MEASUREMENT ERROR," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1477>

This Event is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact [cads@k-state.edu](mailto:cads@k-state.edu).

## **DEVELOPING PREDICTION EQUATIONS FOR FAT FREE LEAN IN THE PRESENCE OF AN UNKNOWN AMOUNT OF PROPORTIONAL MEASUREMENT ERROR**

Zachary Hass<sup>a</sup>, Bruce A. Craig<sup>a</sup>, Allan P. Schinckel<sup>b</sup>

<sup>a</sup> Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN, 47907

<sup>b</sup> Department of Animal Sciences, Purdue University, 915 West State St., West Lafayette, IN, 47907

**Corresponding Author:** Zachary Hass, E-mail address: zhass@purdue.edu

**ABSTRACT:** Published prediction equations for fat-free lean mass are widely used by producers for carcass evaluation. These regression equations are commonly derived under the assumption that the predictors are measured without error. In practice, however, it is known that some predictors, such as backfat and loin muscle depth, are measured imperfectly with variance that is proportional to the mean. Failure to account for these measurement errors will cause bias in the estimated equation. In this paper, we describe an empirical Bayes approach, using technical replicates, to accurately estimate the regression relationship in the presence of proportional measurement error. We demonstrate, via simulation studies, that this Bayesian approach dramatically improves the accuracy of the estimated equation in comparison to the fit from Ordinary Least Squares regression.

**Key words:** Carcass composition, Proportional measurement error, Empirical Bayes

## 1. Introduction

To expedite pork carcass valuation, equations are often used to predict fat-free lean (*FFL*) content from quickly measurable characteristics, such as backfat depth (*BFD*). Of the multiple measuring tools that exist for measuring *BFD*, operator error of the optical probe can lead to readings with error variance that is proportional to the true backfat value (Boland et al., 1995; Schinckel et al., 2010).

When proportional measurement error exists and an analysis is used that assumes the independent variables are measured without error (e.g. Ordinary Least Squares (OLS) regression), it results in a biased regression equation. For example, when the true relationship is linear, this type of measurement error greatly increases the probability of a false positive quadratic term in the prediction equation (Schinckel et al., 2007). Furthermore, the false positive quadratic term is convex, suggesting that the predicted response, such as *FFL* eventually increases as the predictor, *BFD*, increases (Schinckel et al., 2010).

This result has raised concerns about the accuracy of the published equations that have used backfat data collected with an optical probe and feature convex quadratic terms (Johnson et al., 2004; Schinckel et al., 2005; Schinckel et al., 2010; Schinckel, 2012). Previous research introduced an empirical Bayes approach that accurately recovers the true regression coefficients from data with proportional measurement error (Hass et al. 2014). That work, however, assumes prior knowledge of the measurement error proportionality constant  $K$ , which allows the necessary decomposition of the observed variability in *BFD* measurements between population and error-induced variability.

Although such knowledge may reasonably be obtained ahead of time (e.g., a pilot study), we now extend this approach to situations when  $K$  is unknown. This extension is possible provided there are replicate measurements on a subset of pigs. Thus, the objectives of this paper are to propose an empirical Bayes model to estimate carcass composition prediction equations in the presence of proportional measurement error with unknown constant  $K$ , provide a relative size rule regarding the size of the subset of pigs with replicate measurements, and compare results against the equations estimated from OLS regression.

## 2. Background and Methods

### 2.1. Measurement Error

In a typical regression setting, it is assumed that the predictors  $X$  are measured without error. There are times, however, when instead of a true predictor  $X_i$ , we observed  $X_i^*$ , which is a perturbed version of  $X_i$  due to measurement error. The additive measurement model takes the form  $X_i^* = X_i + \delta_i$  and assumes  $E(\delta_i) = 0$  and  $Var(\delta_i) = \sigma_\delta^2$ . The multiplicative (or proportional) measurement error model takes the form  $X_i^* = \delta_i X_i$  and assumes  $E(\delta_i) = 1$  and  $Var(\delta_i) = K$ . The difference between the two models is highlighted by their conditional variances. In the additive case,  $Var(X_i^*|X_i) = \sigma_\delta^2$  and in the multiplicative case,  $Var(X_i^*|X_i) = KX_i^2$ .

When measurement error occurs, it will bias the resulting regression equation. How it biases the regression equation, however, depends on the type of error. Figure 1 illustrates these differences using a simple linear relationship. The black (solid) line represents the true linear model, as well as the average line obtained using OLS (fitting a quadratic model) provided there is no measurement error. The blue (dotted) line represents the average line obtained when using OLS (quadratic model) given additive measurement error. Finally the red (dashed) line represents the average line obtained using OLS (quadratic model) given there is proportional measurement error. The amount of measurement error in both cases is such that the correlation between the true and observed predictor is approximately 0.75.

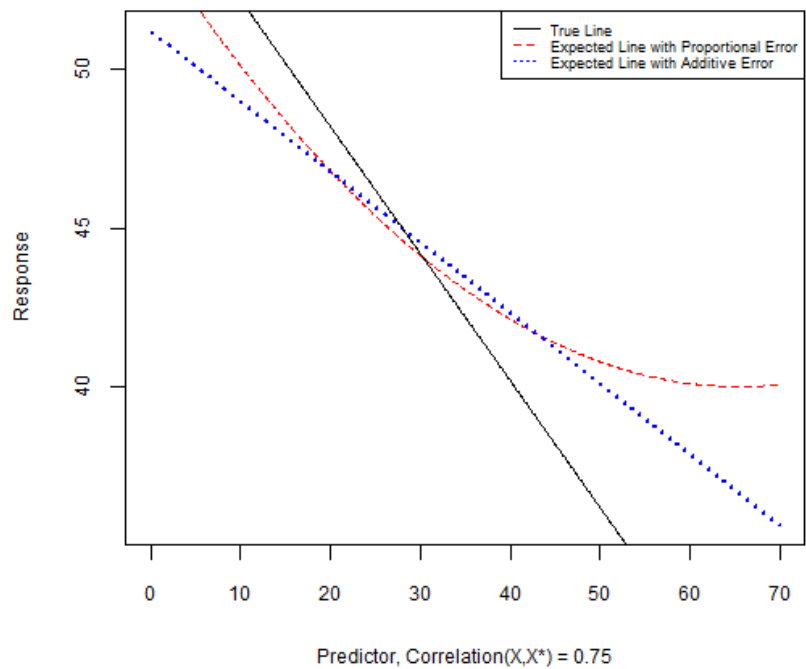
Additive measurement error reduces the perceived linear association between the two variables, thereby reducing the slope of the line towards zero (Carroll et al., 2006). Proportional error, on the other hand, not only reduces the degree of association but also introduces a convex curvature to the relationship (Schinckel et al., 2007). An excellent description of a contrived proportional measurement error problem is given by Hwang (1986).

Several methods have been developed to adjust for additive measurement error but all require an estimate of the measurement error variance. Methods such as SIMEX introduce additional measurement error to extrapolate what the  $\beta$  coefficients would be if there were no error (Carroll, et al., 2006). Others such as regression calibration shrink the observed values  $X^*$  towards the overall mean using the reliability ratio, which is the relative amount of variability in the latent  $X$  over the variability in  $X^*$  (Carroll, et al., 2006).

Extensions of these ideas to the proportional measurement error setting exist. The regression calibration concept can be adapted so that each  $X^*$  has its own reliability ratio as in Hass et al. (2014). Three different ways of approaching the problem with SIMEX are compared through simulation in a discussion paper by Biewen et al (2008). All approaches, however, rely on an estimate, or knowledge of, the measurement error variance. Additionally, estimation of the uncertainty in parameter estimates requires techniques such as the delta method, jackknife, or bootstrap.

We propose a Bayesian approach that treats the  $X$  as an additional unknown. The resulting posterior distribution provides a straightforward method to assess estimation uncertainty in both  $K$  and regression parameters  $\beta$ . Furthermore, provided the experimenter is able to obtain a subset of technical replicates, our model is able to differentiate between a true linear or true quadratic relationship.

**Figure 1: Expected effects of measurement error when fitting OLS**



The black line is the true relationship to be estimated between a predictor  $X$  and some response. The blue line represents the expected quadratic model fit using Ordinary Least Squares (OLS) when the predictor has additive measurement error. The red line represents the expected quadratic model fit using OLS when the predictor has proportional measurement error. The correlation between  $X$  and  $X^*$ , the perturbed predictor, is 0.75 for both types of measurement error.

## 2.2. Statistical Model

For simplicity, suppose we want to estimate the relationship between *BFD* (the predictor,  $X$ ) and *FFL* (the response,  $Y$ ) using a data set of  $N$  pigs. The measured, or observed, *BFD* is designated by  $X^*$  because it is measured with error. Furthermore, we assume that  $R \leq N$  pigs have their *BFD* measured  $S$  times and every other pig has it measured just once.

To estimate the regression relationship between *BFD* and *FFL*, we adopt a Bayesian inference approach. We want to estimate the posterior distribution  $\pi(\beta, \sigma_\epsilon^2, K, X|X^*, Y)$ , where  $\beta$  is the regression coefficients,  $K$  is the unknown proportionality constant, and  $\sigma_\epsilon^2$  is the regression error variance. To account for the measurement error, it is helpful to include the latent variable  $X$ , the Actual Backfat Depth (ABFD), as an additional unknown in the calculations. This inclusion requires specifying a prior for  $X$ .

We consider a Normal prior for  $X$  and use the data to estimate its mean and variance. Technically,  $X$  can only be positive, suggesting a log-Normal or gamma prior. In practice, however, there are very few measurements near zero, suggesting the use of a Normal prior would not be a problem. Because  $E(X^*) = X$  in the usual measurement error context, we estimate the prior mean using  $\hat{\mu}_X = \bar{X}^*$ , the average of the  $N$  pig sample means. For the prior variance  $\sigma_X^2$ , we need to remove the measurement error from the observed variance  $s_{X^*}^2$ . This requires an estimate of  $K$ .

We initially estimate  $K$  from the information in the technical replicates using Equation 1. This equation arises from the fact that the variance of  $X$  is proportional to the mean. Let  $S_{x_i^*}^2$  be the sample variance of the  $i^{\text{th}}$  pig's technical replicate values measured with error. Similarly  $\bar{X}_i^*$  is the  $i^{\text{th}}$  pig's sample mean. If  $R$  pigs are measured  $S$  times then  $\hat{K}$  is the average of the  $R$  ratios of each pig's sample variance and mean. We investigated other estimators of  $K$  and found this estimator to have the lowest variance.

$$\hat{K} = \sum_i \frac{S_{x_i^*}^2}{\bar{X}_i^{*2}} / R; \quad S_{x_i^*}^2 = \sum_1^S \frac{(X_{is}^* - \bar{X}_i^*)^2}{S-1} \quad \text{and} \quad \bar{X}_i^* = \sum_1^S \frac{X_{is}^*}{S} \quad (1)$$

Using  $\hat{K}$  we estimate the prior variance with Equation 2, a method of moments estimator. Note that we now use  $S_{X^*}^2$ , without secondary subscript, to be the sample variance across pig *BFD* measurements. For simplicity, we use only the first replicate measurement from those pigs measured multiple times.

$$\hat{\sigma}_X^2 = \frac{S_{X^*}^2 - \hat{K}\bar{X}^{*2}}{\hat{K}+1}; \quad S_{X^*}^2 = \sum_i^N \frac{(X_{i1}^* - \bar{X}^*)^2}{N-1} \quad \text{and} \quad \bar{X}^* = \sum_i^N \frac{\bar{X}_i^*}{N} \quad (2)$$

For the regression parameters we use Jeffrey's prior  $\pi(\beta, \sigma_\epsilon) \propto \frac{1}{\sigma_\epsilon^2}$  to minimize this prior's influence on the results and "mimic" least squares regression (Kass and Wasserman, 1995). For  $K$  we use an exponential prior with scale parameter  $\lambda = 0.04$  to encompass the likely range of correlations between  $X$  and  $X^*$ . Sensitivity analysis using other  $\lambda$  values did not greatly impact results.

Given these priors, we estimate the desired posterior distribution using Markov chain Monte Carlo (MCMC). We take a Gibbs sampler approach, updating the unknown parameters  $\beta$ ,  $\sigma_\epsilon^2$ ,  $K$ , and  $X$  sequentially while keeping the remaining parameters fixed at their current values. For more details on Gibbs Sampler see Casella and Edwards (1992).

For the regression parameters the updates are sampled directly from their full conditional distributions (see Equations 3 and 4). The  $\chi^2$  stands for the chi-square distribution with degrees of freedom equal to the number of pigs. For the updates of  $X$  and  $K$ , a Metropolis-Hastings (MH) step is used. The conditional distributions are given in Equations 5 and 6. Here and throughout,  $N(\mu, \sigma)$  stands for the Normal distribution with mean and standard deviation. Gamma distributions, designated by *Gam*, are given with shape and scale parameters, respectively. Note that the assumed Gamma distribution for the observed predictor  $X^*$  implies that  $\delta$  follows a Gamma distribution with shape =  $\frac{1}{K}$  and scale =  $K$ . We used a Normal random walk proposal for each MH step, the symmetry of which means only the conditional posterior distributions are needed to form the MH acceptance ratio (Christensen et al., 2011). Subscripts stand for the  $i^{th}$  pig and  $t^{th}$  iteration.

$$\pi(\beta_{t+1}|X_t, Y, \sigma_{\epsilon t}^2) \sim MVN_p \left( (X_t^T X_t)^{-1} X_t^T Y, \sigma_{\epsilon}^2 (X_t^T X_t)^{-1} \right) \quad (3)$$

$$\pi(\sigma_{\epsilon t+1}^2 | \beta_{t+1}, X_t, Y) \sim (Y - X_t^T \beta_{t+1})^T (Y - X_t^T \beta_{t+1}) \frac{1}{\chi^2(N)} \quad (4)$$

$$\pi(X_{it+1} | \beta_{t+1}, Y, \sigma_{\epsilon t+1}^2, X^*) \propto N(X; \mu_X, \sigma_X) N(Y; X_t^T \beta_{t+1}, \sigma_{\epsilon t+1}^2) \text{Gam}(X^*; \frac{1}{K_t}, K_t X_{it}) \quad (5)$$

$$\pi(K_{t+1} | \beta_{t+1}, Y, \sigma_{\epsilon t+1}^2, X^*, X_{t+1}) \propto \prod_i^N \text{Gam} \left( X^*; \frac{1}{K_t}, K_t X_{it+1} \right) \text{Gam} \left( K; 1, \frac{1}{25} \right) \quad (6)$$

We run the Markov chain for 80,000 iterations with a 30,000 iteration burn-in. Posterior means of  $\beta$  and  $\sigma_\epsilon$  were used to describe the estimated prediction equation. Uncertainty in the estimates for a given data set were characterized using 95% credible intervals such that the probability of a parameter being above the interval was equal to the probability of it being below (the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of the posterior distribution).

### 2.3. Simulation Study

To assess performance of the model in recovering the true regression coefficients, we used a simulation study. All data generation, modeling, and analysis was done using R v3.1.1. 1000 data sets were simulated for each of nine scenarios, crossing three levels of measurement error with three experimental unit sizes. Measurement error was set such that the correlation between true and observed  $X$  were approximately 0.75, 0.85, and 0.94 and is controlled by the constant of proportionality designated by  $K$  (0.06, 0.03, 0.01 respectively). These ranges are derived from previous literature (Hass et al., 2014; Schinckel et al., 2007; Schinckel et al., 2010). The number of pigs (experimental unit size) was set to 250, 500, and 1000 pigs to represent a medium study, a large study, and a multi-site study (Johnson et al., 2004).

The true predictor,  $ABFD$ , was sampled from a truncated Normal distribution with a mean of 28 mm and standard deviation of 8 mm with left truncation at 0 (Schinckel et al., 2007). The probability of a negative draw is very small without truncation, but truncation avoids computational and interpretation issues. Measurement error proportional to the mean was introduced by sampling  $BFD$  values for pig  $i$  from a Gamma distribution with mean of  $ABFD_i$  and variance of  $K * ABFD_i^2$ . The response variable,  $FFL$ , was generated from Equation 7 for the primary results and from Equations 8 and 9 for the secondary results. In all three equations the error term ( $\epsilon$ ) came from  $N(0, 3.57)$ . These equations and error standard deviation were derived from previous research (Hass et al., 2014; Johnson et al., 2004; Schinckel et al., 2007).

$$FFL = 56.2 - 0.4 * ABFD + \epsilon \quad (7)$$

$$FFL = 54.46 - 0.543 * ABFD + 0.006 * ABFD^2 + \epsilon \quad (8)$$

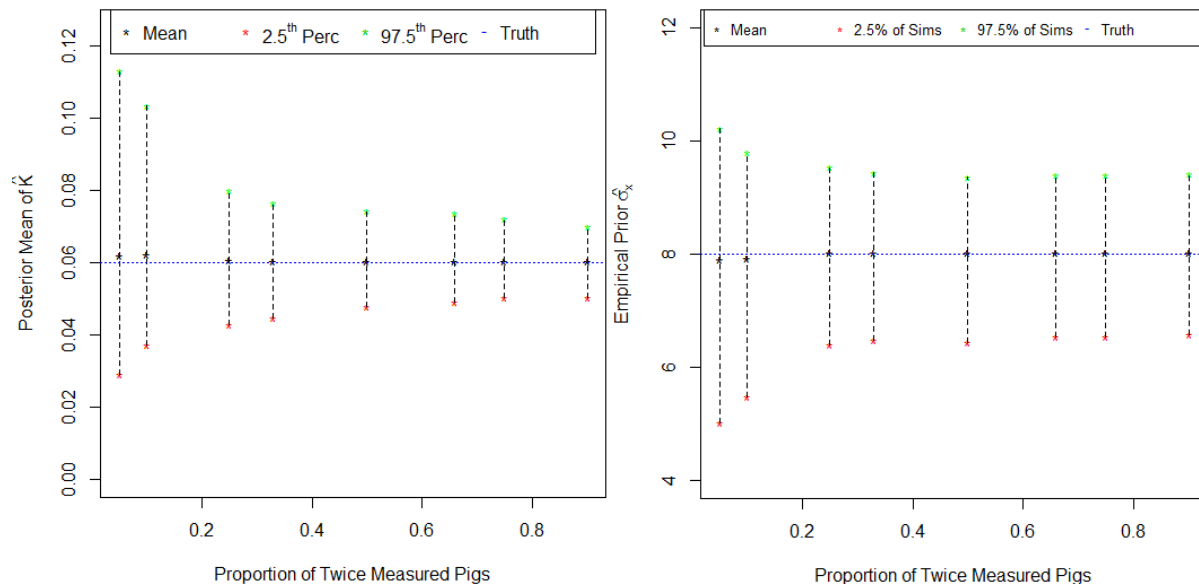
$$FFL = 56.2 - 0.315 * ABFD - 0.003 * ABFD^2 + \epsilon \quad (9)$$

For each simulation, we consider fitting a quadratic relationship between  $BFD$  and  $FFL$  using both our approach and OLS. To examine the performance of our approach, we compare the distribution of posterior means of  $\beta$ ,  $K$ ,  $\sigma_x$ , and  $\sigma_\epsilon$  against the true equation values. For inference about the quadratic term, we use a 95% equal-tailed credible interval for each data set. If 0 falls outside the interval for the quadratic term ( $\beta_2$ ) then we incorrectly infer the presence of a quadratic relationship for that dataset. For OLS, we use the  $P$ -value of the  $t$  test  $H_0: \beta_2 = 0$ . Percentage of false positives (Type I error) across all data sets is compared against the target rate of 5%.

Secondary analyses focus on testing the models' performance when the underlying model is truly quadratic and  $K$  is unknown. Two scenarios were tested. The first scenario generated from a convex quadratic relationship similar to those fit with optical probe data (Johnson et al., 2004) and the second from a concave quadratic relationship where the loss in  $FFL$  is greater per unit increase of  $BFD$  as  $BFD$  increases. Data were generated as before, this time with only 250 pigs and  $K = 0.06$ . The posterior mean (or least-square estimate) of coefficients were saved for each data set and the mean across all data sets were compared graphically.



**Figure 2: Impact of Proportion of Twice Measured Pigs on Parsing Variability in BFD ( $X^*$ )**



The figure on the left describes the distribution of the posterior mean of  $\hat{K}$  across 1000 datasets for different proportion of pigs out of 250 measured twice. The figure on the right does the same for the empirically measured prior  $\hat{\sigma}_x$ . All data were generated using  $K = 0.06$ , a constant of proportionality that implies a correlation of 0.75 between Actual Backfat Depth (*ABFD*) and measured Backfat Depth (*BFD*) depth. The ability to correctly parse observed variability in *BFD* between these two parameters is key to removing the negative effects of proportional measurement error. These plots inform what proportion of pigs must be measured twice.

### 3. Results

#### 3.1. Number of Units Measured Twice

Our first study examined the impact of the proportion of pigs measured twice ( $S=2$ ) on the posterior mean estimates of  $K$  and  $\sigma_x$ . We focus on only two measurements because it represents the fewest additional measurements per pig. The left panel of Figure 2 displays the estimated distribution of the posterior means for both estimators at various proportions of 250 pigs measured twice for 1000 datasets and  $K$  equal to 0.06. As expected, the variability of the  $K$  estimate goes down as the proportion of pigs measured twice increases. However, there are diminishing returns suggesting measuring roughly a third of the pigs twice should give reasonable results. This size is further supported by the right panel in Figure 2, which shows the distribution of the estimate  $\hat{\sigma}_x$  used in the prior for  $X$ . It is nearly unbiased and its variability ceases to decrease noticeably once around a third of the pigs are measured twice. Similar patterns were found for different levels of measurement error and number of pigs. Thus, for the remainder of the simulation study, we consider  $\frac{1}{3}$  of the pigs as measured twice.

#### 3.2. True Linear Relationship

Our second study compares our Bayesian model against least squares regression. Table 1 gives the average estimated parameter values from the Bayesian model, with data generating values presented in the column header. Within each scenario, the model is basically unbiased for the model coefficients, regression variance, the amount of measurement error and the variability in the predictor variable. Correctly parsing apart the sources of variation leads to Type I error rates very near the target of 5%.

Table 2 gives the results of ignoring the measurement error and fitting the least squares regression. Ignoring the measurement error is equivalent to assuming  $K = 0$ , dictating that all of the variability in  $X$  arises naturally. The result is that the estimated standard deviation of  $X$  increases, the coefficients become biased, and Type I error increases dramatically. The Type I error rate worsens as both measurement error and experiment size increase. Bias in the estimated regression line tends to increase as the amount of measurement error increases, within a given sample size.

A visualization of the mean fit between the two methods is given for two scenarios in Figure 3. The plot on the left shows the fit when  $K = 0.06$  and the number of pigs is 250. The plot on the right gives the mean fit for  $K = 0.01$  and 1000 pigs. The larger  $K$  or increased measurement error, leads to greater bias in the coefficients as seen in the increased curvature of the line.

**Table 1: Comparison of parameter estimates across scenarios for Bayesian model**

Number of Pigs	$K$	$\hat{\beta}_0$ $\beta_0 = 56.2$	$\hat{\beta}_1$ $\beta_1 = -0.4$	$\hat{\beta}_2$ $\beta_2 = 0$	Type I Error	$\hat{\sigma}_\epsilon$ $\sigma = 3.57$	$\hat{K}$	$\hat{\sigma}_x$ $\sigma_x = 8$
250	0.01	56.17	-0.395	-0.0001	6.0%	3.56	0.010	7.98
250	0.03	56.29	-0.403	0.0000	5.0%	3.58	0.031	8.01
250	0.06	56.26	-0.397	-0.0002	5.4%	3.57	0.061	8.03
500	0.01	56.12	-0.392	-0.0002	4.3%	3.56	0.010	8.00
500	0.03	56.25	-0.400	0.0000	4.7%	3.57	0.031	8.00
500	0.06	56.28	-0.403	0.0000	4.7%	3.57	0.060	8.05
1000	0.01	56.14	-0.394	-0.0001	3.5%	3.55	0.010	7.99
1000	0.03	56.26	-0.404	0.0001	6.2%	3.56	0.030	8.02
1000	0.06	56.27	-0.405	0.0001	5.9%	3.57	0.060	8.06

Each row represents a scenario of a set number of pigs and amount of measurement error. In each scenario,  $\frac{1}{3}$  of the pigs were measured twice. Each scenario contains 1000 randomly generated data sets. All table entries represent the average posterior mean except for Type I error, which is the proportion of data sets when the equal tailed credible region for  $\hat{\beta}_2$  failed to contain 0. The constant of proportionality ( $K$ ) stands for the amount of measurement error,  $K = 0.01, 0.03,$  and  $0.06$  stands for correlation between actual backfat depth ( $ABFD$ ) ( $X$ ) and measured backfat depth ( $BFD$ ) ( $X^*$ ) of 0.94, 0.85, and 0.75 respectively. The  $\beta$  values are the parameters from the regression model of Fat Free Lean ( $FFL$ ):  $FFL = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ , where  $\epsilon \sim N(0, \sigma_\epsilon)$ , they are estimated by their posterior means. The estimated residual standard deviation is given by  $\hat{\sigma}_\epsilon$  and the estimated standard deviation of  $BFD$  is given by  $\hat{\sigma}_X$ . The former is a posterior mean estimate, while the latter is estimated through method of moments.

**Table 2: Average estimated parameter values from the regression model ignoring measurement error**

Number of Pigs	$K$	$\hat{\beta}_0$ $\beta_0 = 56.2$	$\hat{\beta}_1$ $\beta_1 = -0.4$	$\hat{\beta}_2$ $\beta_2 = 0$	Type I Error	$\hat{\sigma}_\epsilon$ $\sigma_\epsilon = 3.57$	$\hat{\sigma}_x$ $\sigma_x = 8$
250	0.01	56.66	-0.488	0.0024	17.8%	3.72	8.49
250	0.03	55.88	-0.499	0.0035	45.6%	3.92	9.44
250	0.06	54.04	-0.427	0.0033	55.5%	4.11	10.70
500	0.01	56.60	-0.483	0.0022	29.5%	3.73	8.51
500	0.03	55.93	-0.503	0.0036	75.6%	3.93	9.44
500	0.06	54.13	-0.432	0.0033	86.0%	4.12	10.70
1000	0.01	56.62	-0.486	0.0023	53.6%	3.72	8.51
1000	0.03	55.95	-0.504	0.0036	96.6%	3.93	9.45
1000	0.06	54.11	-0.431	0.0033	98.9%	4.11	10.71

Each row represents a scenario of a set number of pigs and amount of measurement error. Each scenario contains 1000 randomly generated data sets. All table entries represent the average estimate except for Type I error, which is the proportion of data sets when the p-value on the quadratic term was less than 0.05. The constant of proportionality ( $K$ ) stands for the amount of measurement error,  $K = 0.01, 0.03,$  and  $0.06$  stands for correlation between Actual Backfat Depth ( $ABFD$ ) ( $X$ ) and measured backfat depth ( $BFD$ ) ( $X^*$ ) of 0.94, 0.85, and 0.75 respectively. The  $\beta$  values are the parameters from the regression model of Fat Free Lean ( $FFL$ );  $FFL = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ , where  $\epsilon \sim N(0, \sigma_\epsilon)$ . The estimated residual standard deviation is given by  $\hat{\sigma}_\epsilon$  and the estimated standard deviation of BFD is given by  $\hat{\sigma}_x$ .

**Figure 3: Comparison of Bayesian and least squares mean fit**

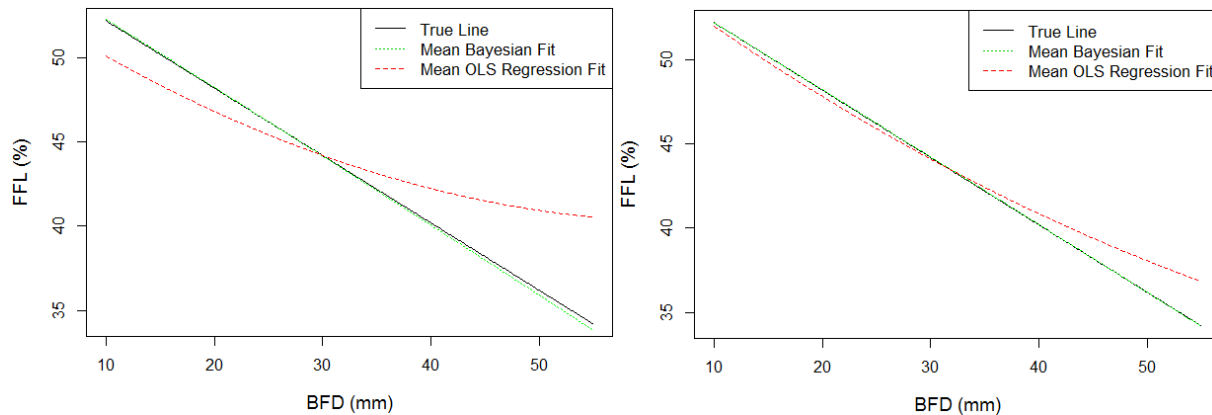
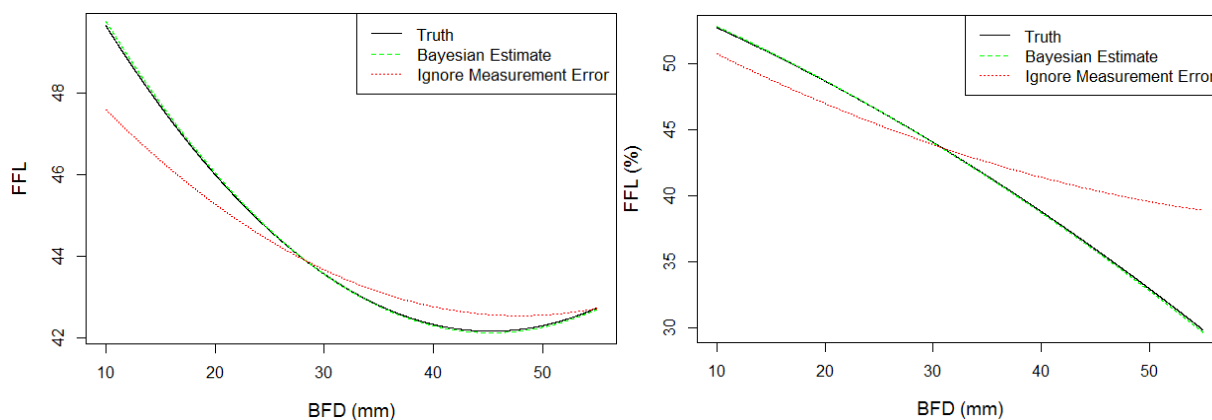


Figure 3 shows the average fit when accounting for measurement error (Bayesian fit) and when ignoring it (OLS fit). In both plots, the black (solid) line is the true relationship. The green (dotted) line is the average Bayesian model fit and the red (dashed) line is the average OLS fit across 1000 data sets. Both methods assume a quadratic relationship between Backfat Depth (*BFD*) and the response, fat free lean (*FFL*). The figure on the left uses datasets of 250 pigs and  $K = 0.06$  (correlation between Actual Backfat Depth (*ABFD*) and measured Backfat Depth (*BFD*) of 0.75) and the one on the right uses datasets of 1000 pigs and  $K = 0.01$  (correlation of 0.94).

**Figure 4: Performance of Bayesian model when data arises from quadratic relationship**



The figure on the left displays the mean estimated relationship across 1000 data sets generated from a convex quadratic relationship.<sup>8</sup> The black line is the true relationship, the green line is the Bayesian model estimate, and the red line is the OLS estimate. The figure on the right is the same, but data were generated from a concave quadratic relationship. Fat free lean (*FFL*) is on the Y-axis. Both figures used datasets with 250 pigs and a constant of proportionality of 0.06 (correlation between true and measured backfat depth of 0.75).

### 3.3. True Quadratic Relationship

All of the results given so far are based on data generated from a true linear relationship. Previous research demonstrated that when  $K$  was known, the Bayesian model accurately estimated the equation when the true relationship was quadratic as well as linear (Hass et al., 2014). To verify that this holds when  $K$  is unknown, we examined two scenarios where the data generating model was quadratic, one with a concave relationship and one with a convex relationship. Results across 1000 data sets are pictured in Fig. 4 and numerically in Table 3. In the convex case ignoring measurement error moderates the relationship and in the concave case ignoring measurement error flips the relationship back to convex. In both cases the Bayesian model accurately recovers the true relationship. Table 3 shows the Bayesian model estimates are nearly unbiased with superior coverage of the quadratic term compared to OLS results.

## 4. Discussion

Ignoring proportional measurement error leads to biased estimation of regression equations that can be very misleading. Without knowledge of  $K$ , it is impossible to know if the observed curvature is due to measurement error or to a truly convex quadratic relationship. If the true relationship is linear, proportional measurement error will introduce a convex quadratic shape (3.2). If the true relationship is quadratic with a convex shape, failing to account for proportional measurement error will lead to a moderated curve. In the case of a true concave shape, the error will flip the curve to convex (3.3).

Published equations have found a convex shape which originally aroused concern that measurement error might be present as discussed in Schinckel et al. (2012). The convex relationship is unlikely, as it would seem implausible that eventually greater  $BFD$  indicates greater lean mass, but the concave relationship, indicating greater loss of lean mass as  $BFD$  increases might be reasonable. In the case of unlikely convex relationships found in the data, it is quite probable that the true relationship is linear or quadratic and concave, but proportional measurement error is bending the curve. Proper use of OLS regression alone will not detect or correct for this problem.

If proportional measurement error is suspected, some correction must be made to avoid biased results. Our Bayesian model is effective for correcting the bias, provided there is knowledge of the amount of variability in the predictor due to the measurement error. This knowledge can be expressed either through prior knowledge of the variability in  $X$  or in the correlation between true and observed values in the measurement process or both. This would take the practical form of directly specifying  $\sigma_X$  or  $K$  in the model (2.2) as in Hass et al. (2014). In the absence of such knowledge, collecting technical replicates by measuring as few as a third of pigs twice is sufficient to be nearly unbiased for both the amount of measurement error and the true relationship between  $X$  and  $Y$  (3.1), which is the extension presented in this paper.

This paper focused on just two measurements per selected pig because it represents the fewest number of additional measurements per pig. Additional technical replicates per pig ( $S > 2$ ) will lead to a more precise estimate of  $K$ , but whether this increase in precision is worth the

**Table 3: Comparing Bayesian model and least squares for quadratic data**

Estimation Method	$\hat{\beta}_0$ $\beta_0 = 54.46$	$\hat{\beta}_1$ $\beta_1 = -0.543$	$\hat{\beta}_2$ $\beta_2 = 0.0060$	Coverage	$\hat{\sigma}_\epsilon$ $\sigma_\epsilon = 3.57$	$\hat{K}$ $K = 0.06$	$\hat{\sigma}_x$ $\sigma_x = 8$
Least Squares	50.58	-0.335	0.0035	55.2%	3.72	-	10.69
Bayesian Model	54.66	-0.553	0.0061	97.4%	3.54	0.061	8.03
	$\hat{\beta}_0$ $\beta_0 = 56.2$	$\hat{\beta}_1$ $\beta_1 = -0.315$	$\hat{\beta}_2$ $\beta_2 = -0.003$				
Least Squares	55.16	-0.472	0.0032	4.7%	4.36	-	10.69
Bayesian Model	56.35	-0.321	-0.0030	93.4%	3.56	0.060	8.05

Each row represents a scenario of 250 pigs,  $\frac{1}{3}$  of them measured twice and proportionality constant ( $K$ ) of 0.06 (correlation between Actual Backfat Depth ( $ABFD$ ) ( $X$ ) and measured Backfat Depth ( $BFD$ ) ( $X^*$ ) of 0.75). Each scenario contains 1000 randomly generated data sets. The table entries are the average estimate except for coverage, which is the proportion of data sets when the procedure correctly inferred that  $\beta_2 = 0$ . The  $\beta$  values are the parameters from the regression model of Fat Free Lean (FFL);  $FFL = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ , where  $\epsilon \sim N(0, \sigma_\epsilon)$ . The estimated residual standard deviation is given by  $\hat{\sigma}_\epsilon$  and the estimated standard deviation of BFD is given by  $\hat{\sigma}_x$ . There were two underlying regression relationships tested, the first three rows refer to a convex quadratic relationship, and the last three rows refer to a concave quadratic relationship. Least squares refers to Ordinary Least Squares (OLS) regression estimates that ignore the presence of measurement error while the Bayesian model takes the measurement error into account.

additional cost and time in collection is debatable given how well this method worked with  $S=2$ . If multiple measurements per pig are not an issue, alternative measurement designs that focus on more technical replicates on a smaller proportion of pigs may prove beneficial.

Our Bayesian approach very accurately estimates the equation regardless of whether the true relationship between  $X$  and  $Y$  is linear or quadratic. This approach naturally provides measures of uncertainty for any of the sampled quantities including  $\beta$  and  $K$ . Alternative estimators of  $K$  exist although not discussed in this work. Prediction equations often feature additional predictors and this model extends easily for the inclusion of those predictors measured without error. The motivating example for this work is the prediction of pork carcass composition, but the framework should generalize more broadly to any carcass evaluation using a measurement containing proportional measurement errors.

It should be noted that a danger of the method of moment variance estimator in Equation 2 is the possibility of a negative estimate. This is possible if the correlation between true value and measurement degrade considerably (around 0.45 in our setting) without inflating the variance of the observed values above the variance of the true measurements. Since this is a clear violation of the assumptions of our model, we advise the reader who encounters a negative estimate in Equation 2, to adjust the modeling assumptions or look into alternative solutions.

Our largest simulation study size of 1000, representing the size of previous studies that have combined multiple sites did not simulate data as if from multiple studies. If the reader has data with such a structure, it is necessary to model the potential correlation within a site.

In addition to valuation, prediction equations for carcass lean mass are also used commercially by pork producers, for teaching, research, and Extension. They play important roles at swine shows and for 4-H livestock projects. Estimated *FFL* is used to predict the dietary lysine requirements of grow-finish pigs. Bias in the equations used for that purpose has been estimated, in one case, to decrease profitability by \$5.40 per pig (Schinckel et al., 2012). Therefore, it is important for economic as well as educational purposes that these equations be accurate. This can only be accomplished if measurement error is properly accounted for.

All results are based on simulation study and therefore do not speak to any practical issues that may arise when working with data collected from an actual live animal experiment. Particularly the difficulties of taking multiple measurements on a single carcass with an evasive instrument. Measurements of loin muscle depth are similarly prone to measurement error and are correlated with *BFD* as discussed in Olsen et al (2007) and Schinckel et al (2010). The focus of future work should be on the adaptation of this model to handle such data.

## Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.



## References

- Boland, M. A., Berg, E. P., Akridge, J. T., & Forrest, J. C. (1995). The impact of operator error using optical probes to estimate pork carcass value. *Review of Agricultural Economics*, 17(2), 193-204.
- Biewen, E., & Rosemann, M. (2008). Multiplicative Measurement Error and the Simulation Extrapolation Method. *Sandra and Rosemann, Martin, Multiplicative Measurement Error and the Simulation Extrapolation Method (January 1, 2008)*.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.
- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3), 167-174.
- Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press.
- Hass, Z., Zhou, B.A. Craig. (2014). Developing Prediction Equations for Carcass Lean Mass in the Presence of Proportional Measurement Error. *Conference on Applied Statistics in Agriculture Proceedings*. 26, 115-129.
- Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the US Department of Energy. *Journal of the American Statistical Association*, 81(395), 680-688.
- Johnson, R. K., Berg, E. P., Goodwin, R., Mabry, J. W., Miller, R. K., Robison, O. W., ... & Tokach, M. D. (2004). Evaluation of procedures to predict fat-free lean in swine carcasses. *Journal of animal science*, 82(8), 2428-2441.
- Kass, R.E., and L. Wasserman. 1995. A Short Course on Applied Bayesian Statistics. *Self Published*, Pittsburgh, PA. 17-18.
- Olsen, E. V., Candek-Potokar, M., Oksama, M., Kien, S., Lisiak, D., & Busk, H. (2007). On-line measurements in pig carcass classification: Repeatability and variation caused by the operator and the copy of instrument. *Meat science*, 75(1), 29-38.

Schinckel, A. P. (2005). Critique of “Evaluation of procedures to predict fat-free lean in swine carcasses”. *Journal of animal science*, 83(12), 2719-2720.

Schinckel, A. P., Einstein, M. E., Foster, K., & Craig, B. A. (2007). Evaluation of the impact of errors in the measurement of backfat depth on the prediction of fat-free lean mass. *Journal of animal science*, 85(8), 2031-2042.

Schinckel, A. P., Wagner, J. R., Forrest, J. C., & Einstein, M. E. (2010). Evaluation of the prediction of alternative measures of pork carcass composition by three optical probes. *Journal of animal science*, 88(2), 767-794.

Schinckel, A. P., & Rusk, C. P. (2012). The need for accurate prediction equations for the carcass lean content of pigs. *Journal of Extension*, 50(3).