Assignment 9

Sam Olson

1.

Define random variables and covariates appropriate to develop a regression model to relate OXY to bm. Examine the scatterplot of OXY on bm. Comment on features of these data based on visual examination of the scatterplot. In particular, identify any characteristics that should be accommodated by a random component for this problem.

Answer

Oxy is a random variable, and the covariate is body mass. Oxy is a positive valued quantity, with minimum 1.27 and maximum 121.67, with units "ng/g wet wt."; furthermore, we have 109 observations in the dataset.

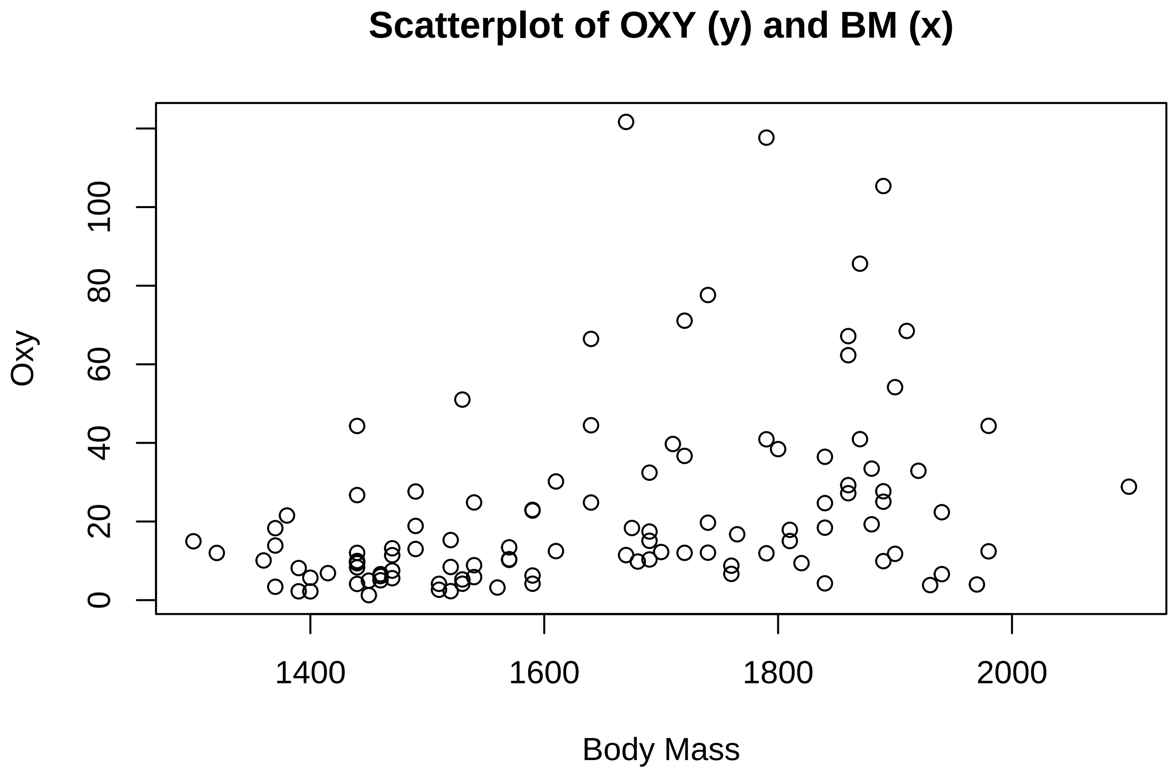
Given this information, we define 109 random variables as $\{Y_i : i = 1, 2, ..., 109\}$, and let Y_i denote a random variable for Oxy (ng/g wet wt.) for the i-th glaucous gull. Further, let $x_i = (1,\ \text{bm}_i)^\top$ with bm_i denote the body mass covariate for the i-th glaucous gull. Because Oxy is strictly positive, we can take the support $Y_i \in (0, \infty)$.

Using the GLM structure for the regression model:

$$\eta_i = x_i^\top \beta, \quad g(\mu_i) = \eta_i \Rightarrow g\big(E[Y_i]\big) = x_i^\top \beta$$

With Y_i from an exponential-dispersion family (random component), η_i the linear predictor, and $g(\cdot)$ the link (systematic component).

Now, regarding (potential) random component(s):



From the scatterplot, we have some sense that variability increases for larger values of Body Mass; the initial thought being this *could* rule out a Normal random component which assumes a constant variance function. We will need to dig deeper in Question 2 however, as it is perhaps not as obvious whether a Gamma or an Inverse Gaussian is a better random component candidate (the remaining possible random components, given the response random variable is continuous and not discrete, which rules out Poisson, Binary, and Binomial random components). There is also evidence of right-skewness in Oxy.

2.

Examine the issue of random model component choice more closely, using approaches we discussed in class.

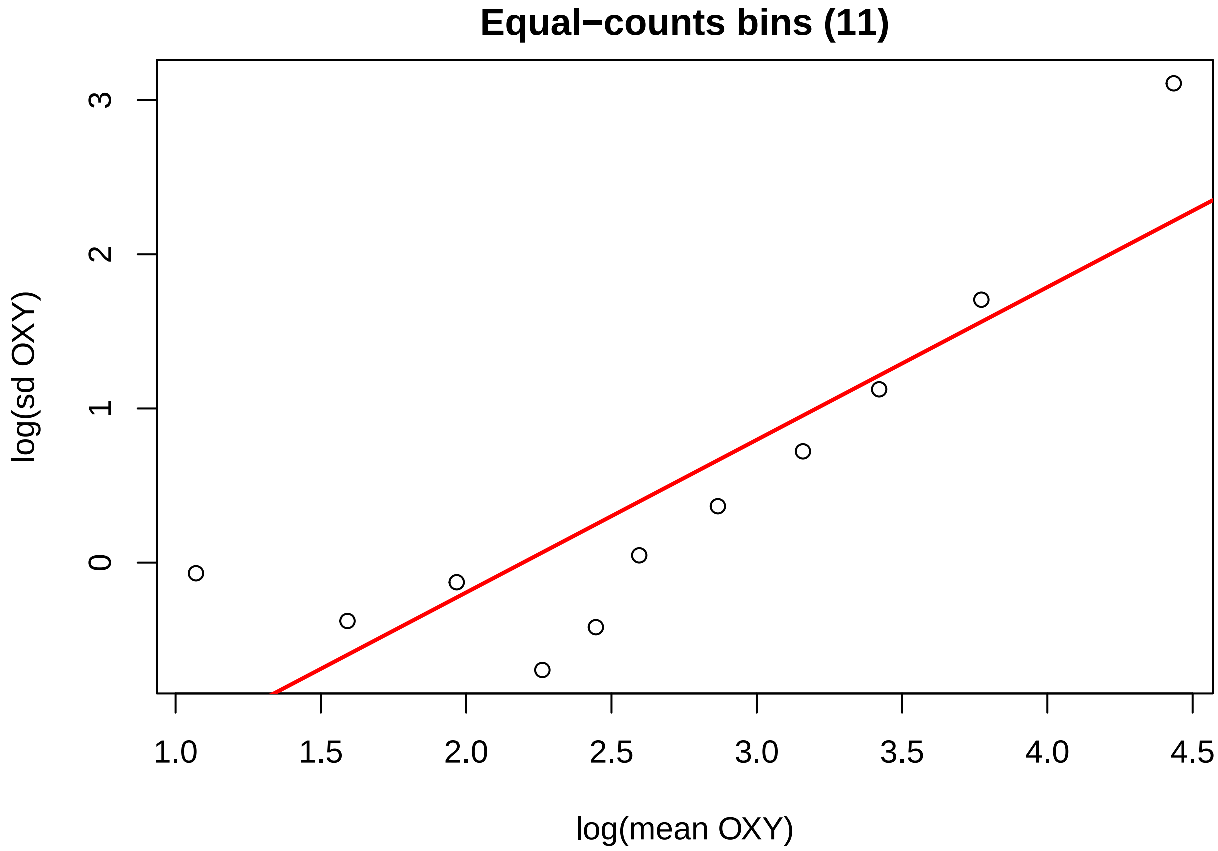
NOTE: It may very well be the case that there are two potential random components that are difficult to distinguish between at this point.

Answer

We'll first consider a Box-Cox Plot to evaluate the relationship between log(mean(Oxy)) and log(sd(Oxy)). We will start with an arbitrary binning into 11 bins, roughly 10 obs per bin, and also evaluate whether the results differ when considering other binning procedures

Table 1: Box–Cox mean–sd regression for 11 equal-count bins: log(sd) log(mean)

	Bins	Term	Estimate	SE	t_value	p_value	R2	Adj_R2	N
(Intercept)	11	(Intercept)	-2.1748	0.5534	-3.93	3.46e-03	0.743	0.714	12
log(m)	11	log(m)	0.9905	0.1944	5.10	6.49e-04	0.743	0.714	12

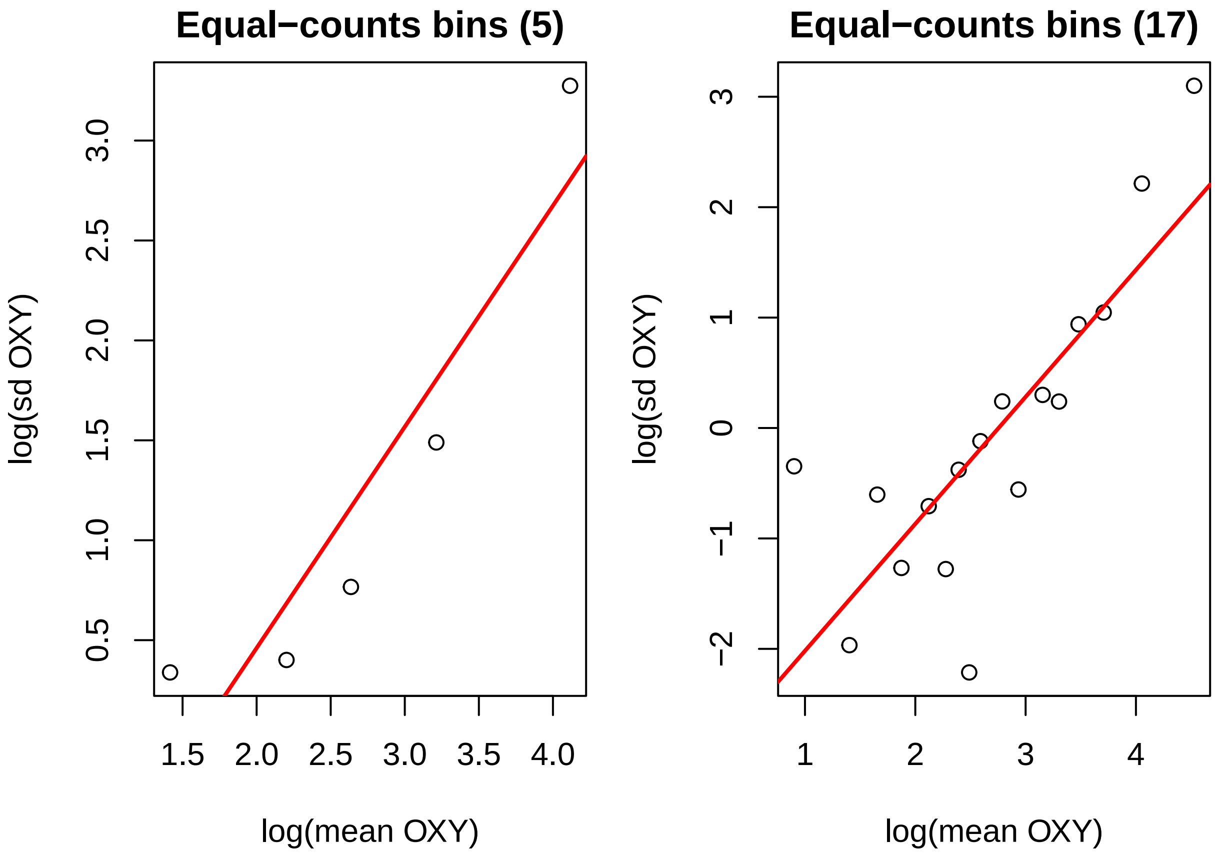


With $V(Y) \propto \mu^{\theta}$, and for $\theta = 2 * \text{Slope} = 1.98 \rightarrow V(Y) \propto \mu^2$.

Now we have some evidence now to support a Gamma random component for μ^2. As an extra validation though, let's consider some of binning(s) to ensure this isn't an artefact of our particular binning method used for previous Box-Cox plot.

Table 2: Linear regressions for Box–Cox mean–sd plots (log(sd) log(mean))

	Bins	Term	Estimate	SE	t_value	p_value	R2	Adj_R2	N
(Intercept)	5	(Intercept)	-1.7530	0.7405	-2.37	9.88e-02	0.859	0.813	6
log(m)	5	log(m)	1.1069	0.2584	4.28	2.34e-02	0.859	0.813	6
(Intercept)1	17	(Intercept)	-3.1680	0.6184	-5.12	1.25e-04	0.651	0.627	18
log(m)1	17	log(m)	1.1500	0.2176	5.29	9.14e-05	0.651	0.627	18



We can still reasonably justify the Gamma random component, but to safeguard against being wrong we'll still consider the possibility of an inverse Gaussian, as we do still have visual evidence to support variance being related to the expectation.

3.

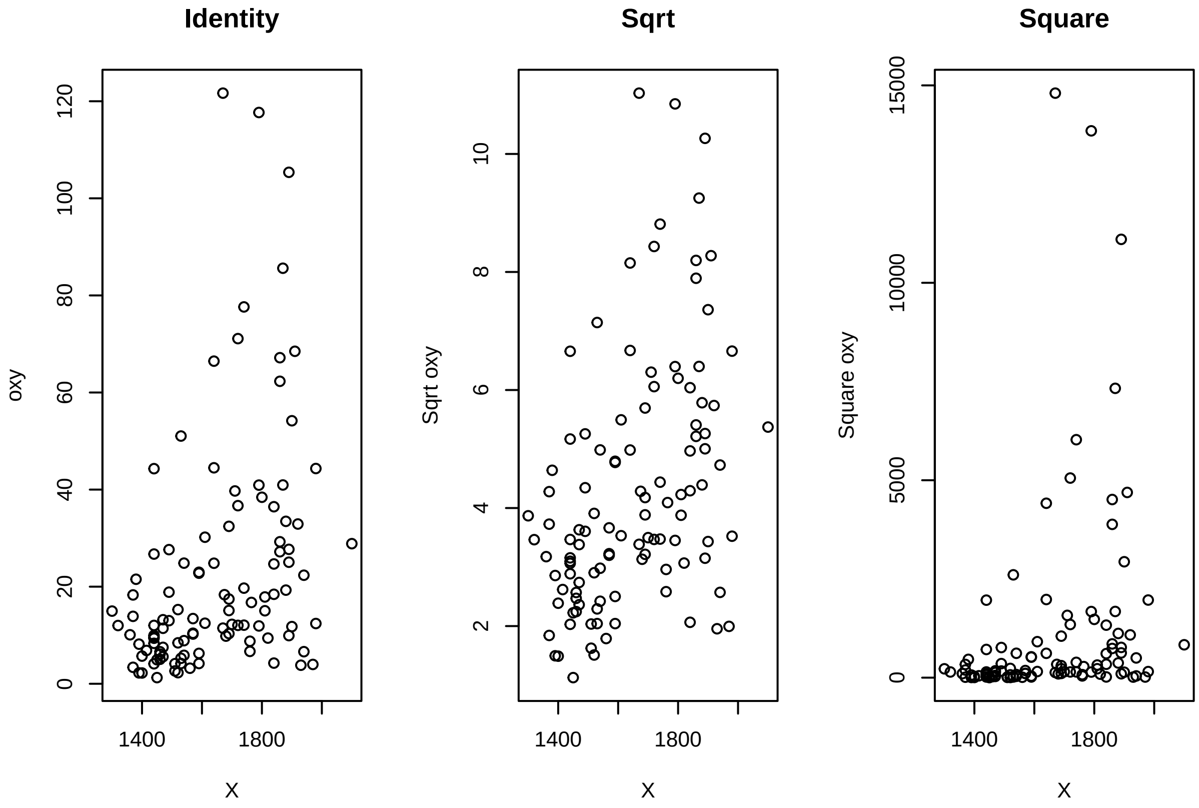
Suggest what you believe is a good link function for the problem. Present supporting evidence for your choice. Again, it may be difficult to make a clear choice between possibilities.

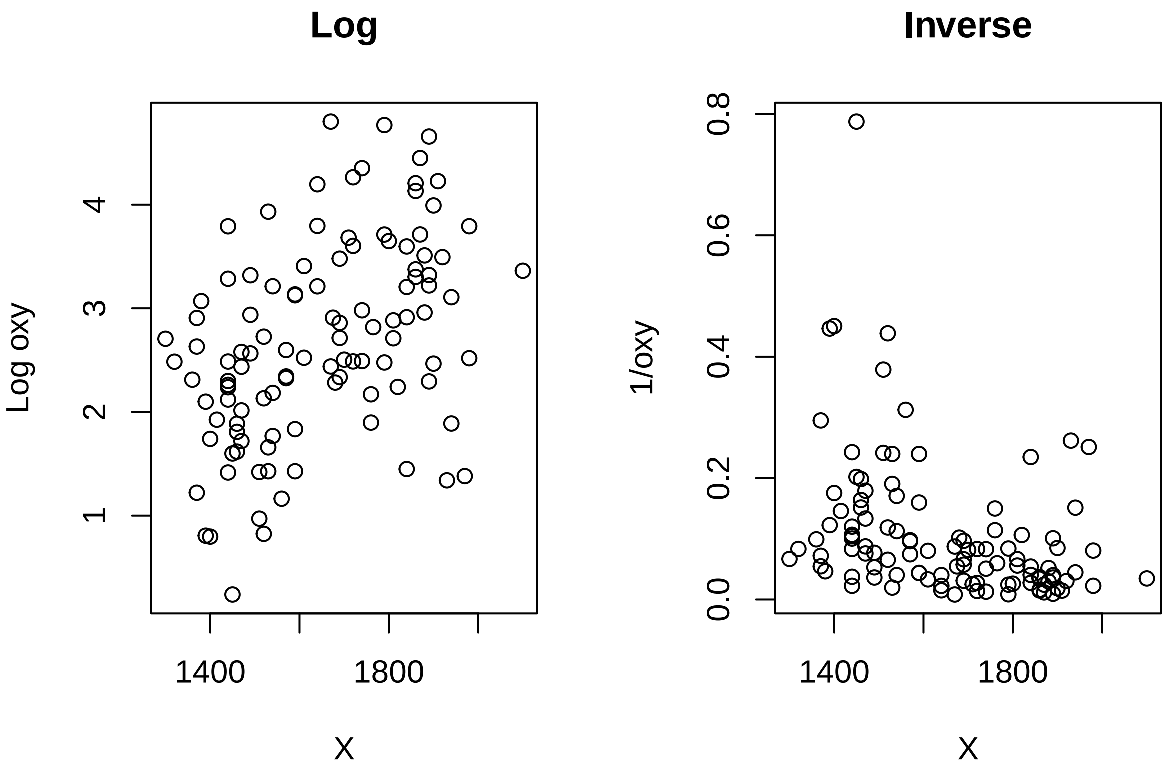
Answer

To consider appropriate link functions, we're looking for a transformation T, such that $y_i = T(x_i\beta)$. So let's start by considering a few possible transformations:

- log-link (log transformation)
- sqrt-link (sqrt transformation)
- power-link (square transformation)
- identity link (no transformation)

Generally, we're looking for a linear relationship between X and Y when doing this transformation, for the purpose of identifying an appropriate link.





None of these transformations look especially great in terms of linear fit, though, that being said, the log link doesn't seem especially bad (perhaps a 'least worst' among the transformations considered).

4.

Fit models with up to two different random components, but using a log link function for both. Estimate regression parameters using maximum likelihood, combined with the usual moment-based estimate of ϕ. Compute Wald theory intervals for the elements of β, unscaled and scaled deviances, and maximized log likelihoods.

Answer

Note: The modelling was done using Kaiser's `basic.glm` function, such that deviance residuals were easier to extract than using the typical `glm` function (with it's wonky Fisher-iteration residuals, or whatever they are.) Also, after confirming with Kaiser, the below intervals are 95% intervals (they just need to be specified by the user in this homework, allegedly).

Table 3: Model comparison with Wald 95% CIs and both unscaled and scaled deviances

Term	Estimate	SE	Phi	UnscaledDev	ScaledDev	LogLik	LCL	UCL	Model
Intercept	-0.9606	0.7734	1.1673	76.5991	89.4139	-144.0595	-2.4765	0.5553	Gamma (log link)
Body Mass	0.0024	0.0005	1.1673	76.5991	89.4139	-144.0595	0.0015	0.0033	Gamma (log link)
Intercept	-1.2291	0.7843	23.2506	5.7556	133.8205	104.5647	-2.7664	0.3082	Inverse Gaussian (log link)
Body Mass	0.0026	0.0005	23.2506	5.7556	133.8205	104.5647	0.0016	0.0036	Inverse Gaussian (log link)

5.

The estimates of ϕ will be quite different between your two models but this is to be expected because of the different distributional forms involved. To see how the models are reflecting variances, compute the variance for a response distribution at several values of the covariate.

Answer

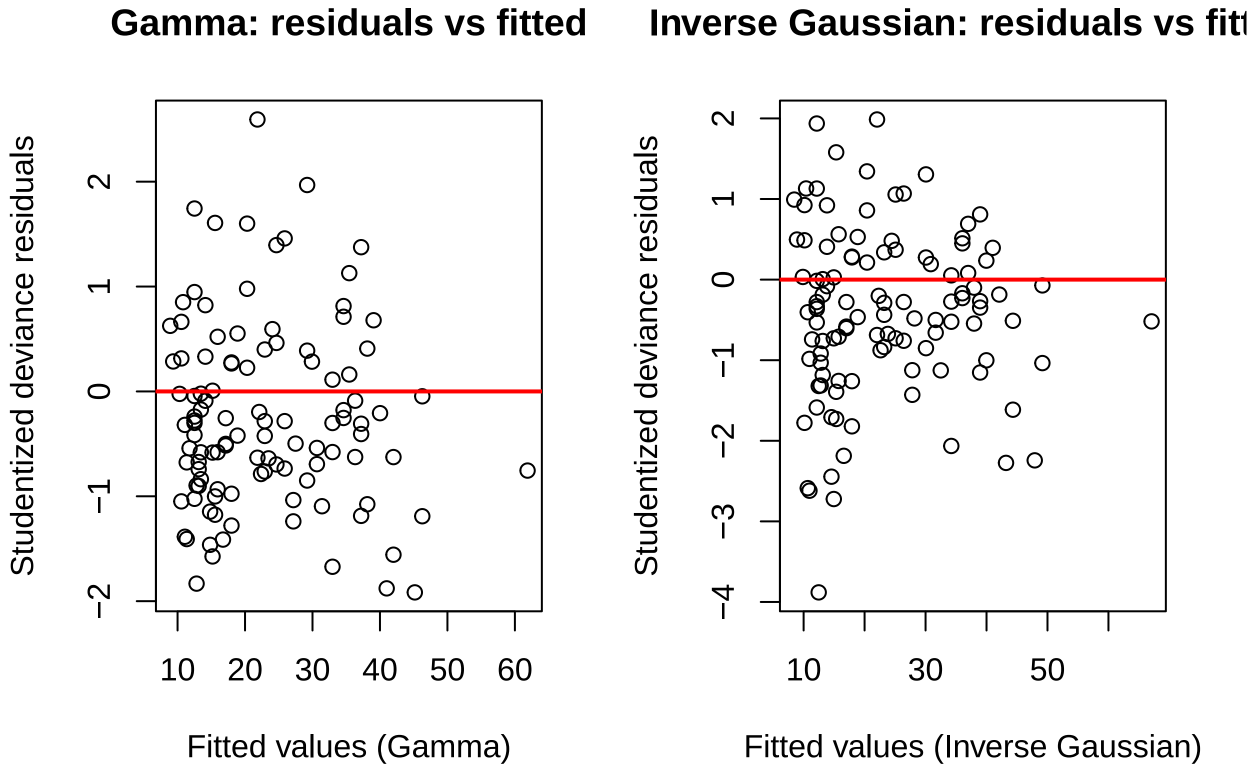
	bm	mu_gamma	mu_ig	var_gamma	var_ig
0%	1300	8.915	8.460	92.7710	14077.30
25%	1470	13.456	13.135	211.3551	52691.95
50%	1640	20.310	20.394	481.5190	197228.21
75%	1820	31.408	32.496	1151.4614	797834.67
100%	2100	61.877	67.071	4469.2914	7015325.31

Sure enough, the variance of the Inverse Gaussian random component GLM is dramatically larger than the Gamma random component GLM, though as perhaps expected their μ_i values are very close to one another (though if expectation were all we were interested in, we'd probably want to consider something other than a GLM for modelling).

6.

Produce studentized deviance residual plots (residuals versus fitted values) for the two random components you are investigating. Do these assist you in distinguishing between the two possible models?

Answer



Gamma looks a bit better, i.e., the residual spread is more constant across fitted values under Gamma; by contrast, the Inverse Gaussian has some pattern in having larger negative studentized deviance residuals.

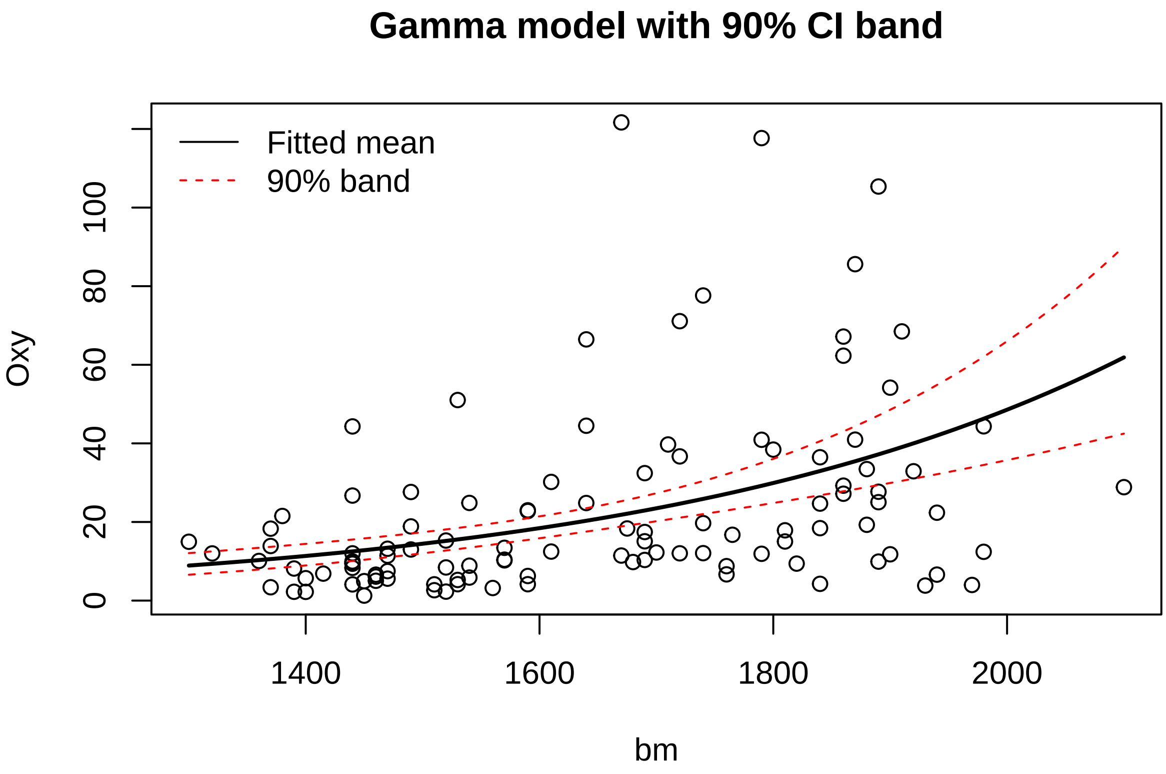
7.

Pick one of your two models, compute Wald theory intervals for the regression parameters and produce a pointwise 90% confidence band for the regression function.

Answer

Table 5: Wald 90% Confidence Intervals (Gamma model)

Parameter	Estimate	SE	LCL	UCL
β_0 (Intercept)	-0.961	0.7734	-2.233	0.312
β_1 (Body Mass)	0.002	0.0005	0.002	0.003


Gamma model with 90% CI band
Fitted mean
90% band
Oxy
bm