Assessing the Impact of Noise on Regression Coefficients in Predictive Models: A Study Using ACTG175 Data

Introduction and Topic Description

In predictive modeling, noise in the data can significantly impact the reliability and accuracy of regression results. Noise refers to random variations or errors in the predictor variables, which can lead to biased or unstable estimates of the regression coefficients. This is especially concerning in real-world datasets, where noise is often unavoidable. The presence of noise can distort the true relationship between the predictor variables and the outcome, resulting in incorrect inferences and predictions.

This study uses the ACTG175 dataset, which includes clinical data on HIV treatment, to investigate how different levels of noise affect regression coefficient estimates and r-squared. By introducing noise into key predictors—such as age, weight, karnofsky score, and CD4 count—the study examines how noise influences the stability and accuracy of regression models. In particular, we explore how noisy estimates compare to the true coefficients and assess correction methods to mitigate bias caused by noise. Through visualizations, we demonstrate the effects of varying noise levels on model performance and discuss how correction formulas can be applied to obtain more accurate and unbiased estimates. Standard errors are also calculated to assess the precision of the unbiased estimates.

Data Description

The ACTG175 dataset, from the `speff2trial` package in R, includes demographic and clinical variables from an HIV treatment trial. Key variables selected for this study are:

- **cd420**: Viral load (log scale).
- **age**: Patient's age in years.
- **wtkg**: Weight in kilograms.
- **karnof**: Karnofsky performance score.
- **cd40**: Measure of disease progression (CD4 count).

There are no missing values in these variables, and the dataset is pre-processed by selecting the relevant columns for analysis.

Methodology

The analysis assesses the effect of noise on regression coefficients, r-squared, and explores unbiased correction methods for estimates and standard errors. Visualizations are provided to assess the impact of noise.

Modeling Approach

A linear regression model predicts `cd420` using predictors `age`, `wtkg`, `karnof`, and `cd40`. The model is first fitted to the noise-free data to establish baseline coefficients. The regression model for the noise-free case is:

$$\text{cd420} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{wtkg} + \beta_3 \cdot \text{karnof} + \beta_4 \cdot \text{cd40} + \epsilon$$

where ϵ represents the residual error.

Noise is added to the predictors by introducing normally distributed random noise with varying standard deviations to simulate measurement errors. The noise levels used are 0.1, 0.5, 1.0, 2.0, 5.0 . For each noise level, random noise $\epsilon_{\text{noise}} \sim \mathcal{N}(0, \sigma^2)$ is added to the predictors, where σ is the standard deviation corresponding to each noise level.

The regression model for the noisy data is:

$$\text{cd420} = \beta_0 + \beta_1 \cdot (\text{age} + \epsilon_{\text{age}}) + \beta_2 \cdot (\text{wtkg} + \epsilon_{\text{wtkg}}) + \beta_3 \cdot (\text{karnof} + \epsilon_{\text{karnof}}) + \beta_4 \cdot (\text{cd40} + \epsilon_{\text{cd40}}) + \epsilon_{\text{noise}}$$

where $\epsilon_{\text{age}}, \epsilon_{\text{wtkg}}, \epsilon_{\text{karnof}}, \epsilon_{\text{cd40}}$ represent the noise added to each predictor, and ϵ_{noise} is the residual error in the noisy model.

Unbiased Estimates

A correction formula adjusts the predictor matrix (X) to account for noise variance, providing unbiased coefficient estimates. These corrected estimates are compared to both noisy and true coefficients. The corrected coefficient estimate $\hat{\beta}$ is given by the following formula:

$$\hat{\beta} = \left(\frac{X^T X}{n} - S^2 \right)^{-1} \frac{X^T y}{n}$$

where:

- $\hat{\beta}$ is the corrected estimate, X is the matrix of predictors, S^2 is the noise variance/noise level, n is the sample size, and y is the vector of observations.

This approach is statistically consistent, as demonstrated by Evans and King (2023) .

Standard Error Calculation

The standard error of the regression coefficients is calculated using the variance-covariance matrix of the estimated coefficients. The variance-covariance matrix is derived from the formula:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

where σ^2 is the estimated error variance, and $(X^T X)^{-1}$ is the inverse of the matrix product of the predictor matrix X and its transpose. The standard errors for each coefficient are the square roots of the diagonal elements of this variance-covariance matrix. Also the noise level was also subtracted to get acurate estimates.

Statistical Evaluation

Regression model performance is evaluated based on:

- **Coefficient Estimates**: Comparison between noisy estimates and true coefficients.
- **R-Squared Values**: Comparison of R-squared values for noisy and corrected models.
- **Standard Errors**: Comparison of the standard errors calculated from noisy and corrected models.

To prevent the intercept from distorting the graph due to its large value and resulting skewness, it will be omitted from the graph and certain tables.

Results

True Coefficients and Model Fit

The true regression coefficients, R^2, and standard errors (SE) for the noise-free data are summarized below:

Metric	Value
Intercept (Coefficient)	11.50884
age (Coefficient)	-0.26246
wtkg (Coefficient)	0.00249
karnof (Coefficient)	1.27151
cd40 (Coefficient)	0.70613
R^2	0.34364
Intercept (SE)	45.17220
age (SE)	0.29573
wtkg (SE)	0.19333
karnof (SE)	0.43365
cd40 (SE)	0.02149

Table 1: True regression coefficients, R^2, and standard errors.

Noisy Coefficients

Key Observations

- The intercept shows the most variability, increasing significantly as noise levels rise.
- Coefficients for predictors (`age`, `wtkg`, `karnof`, and `cd40`) remain relatively stable at lower noise levels but deviate at higher noise levels. However we can see some estimates like age increase significatly at 2.0 and decreases at 5.0.
- The R^2 values remain robust across different noise levels, showing minimal changes. The table will be displayed below:

Noise SD	Intercept	Age Coef.	Wtkg Coef.	Karnof Coef.	Cd40 Coef.	R^2
0.1	11.88034	-0.26291	0.00144	1.26842	0.70619	0.34364
0.5	12.95368	-0.25653	-0.00174	1.25690	0.70630	0.34355
1.0	20.77076	-0.25210	-0.00641	1.17684	0.70638	0.34340
2.0	14.93909	-0.19742	-0.03078	1.23504	0.70681	0.34386
5.0	67.78933	-0.14872	-0.07570	0.70105	0.70595	0.34036

Table 2: Noisy regression coefficients and R^2 values under varying noise levels.

Unbiased (Corrected) Estimates

The corrected estimates show better stability at higher noise levels, with r-squared values closer to the true value. However, they deviate more at lower noise levels, where noisy estimates sometimes outperform them. This suggests the correction method may overcompensate under low-noise conditions, introducing bias. Notably, the cd40 coefficient remains stable across most noise levels, except at 5.0, indicating varying sensitivity across variables.

While the correction process improves overall stability, its limitations in addressing bias for certain variables highlight the need for refinement, such as adaptive corrections tailored to noise levels.

Noise SD	**Age Coef.**	**Wtkg Coef.**	**Karnof Coef.**	**Cd40 Coef.**	**R^2 Corrected**
0.1	-0.23300	0.01650	1.37000	0.70700	0.34369
0.5	-0.23800	0.03020	1.36000	0.70700	0.34364
1.0	-0.27200	0.00561	1.40000	0.70500	0.34406
2.0	-0.30300	-0.01650	1.42000	0.70600	0.34349
5.0	-0.36900	-0.08520	1.52000	0.70200	0.34025

Table 3: Corrected (Unbiased) Regression Coefficients and R^2 for Varying Noise Levels.

Noisy and Corrected Estimates with Standard Errors

Below are the noisy and unbiased standard error estimates at each noise level. The noisy estimates are presented first, followed by the corrected estimates. These values reflect the impact of noise and the correction process on the coefficients and their precision.

Noise SD	**Age SE**	**Wtkg SE**	**Karnof SE**	**Cd40 SE**	**Type**
0.1	0.29570	0.19334	0.43367	0.02149	Noisy
0.5	0.29528	0.19303	0.43249	0.02148	Noisy
1.0	0.29477	0.19275	0.42752	0.02150	Noisy
2.0	0.28600	0.19138	0.41164	0.02148	Noisy
5.0	0.25746	0.18042	0.32814	0.02149	Noisy
0.1	0.28327	0.18749	0.17648	0.02138	Corrected
0.5	0.28303	0.18734	0.17649	0.02137	Corrected
1.0	0.28020	0.18678	0.17541	0.02138	Corrected
2.0	0.27659	0.18494	0.17255	0.02131	Corrected
5.0	0.25083	0.17135	0.16115	0.02122	Corrected

Table 4: Noisy and Corrected Regression Coefficients Standard Errors for Varying Noise Levels.

The corrected estimates consistently show lower values across all four measurements, indicating a systematic adjustment that reduces variability. However, they deviate from the true value as noise increases. In contrast, the noisy estimates exhibit greater variability in standard errors, particularly for the age and karnof coefficients, suggesting lower precision at higher noise levels, despite being closer to the true value. Interestingly, for age, the standard error decreases as noise increases, potentially due to the model shrinking the coefficient estimates toward zero. Meanwhile, the standard errors for the corrected estimates remain consistently lower as noise increases, highlighting the effectiveness of the bias correction in reducing variability.

Visualization of Plots

Figure 1: This plot shows the true, noisy, and corrected estimates for predictors age, weight, karnof, and cd40, emphasizing the corrected estimates' stability over the noisy ones.
Figure 2: This plot compares R-squared values of noisy and corrected models across varying noise levels, showing a slight decline in corrected estimates as noise increases.
Figure 3: This plot compares the standard errors of noisy and corrected estimates for predictors, highlighting reduced variability with noise correction.

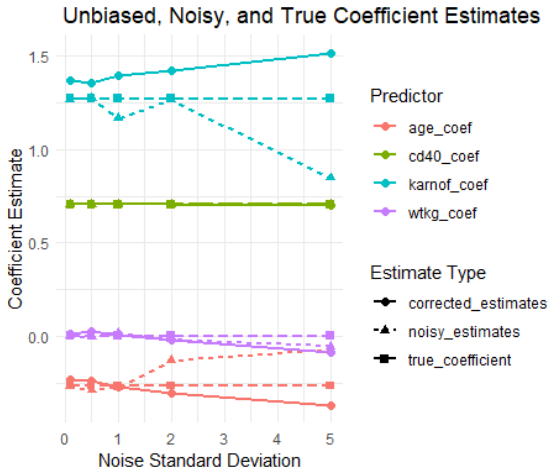

Figure 1: True, Noisy and Corrected Estimates

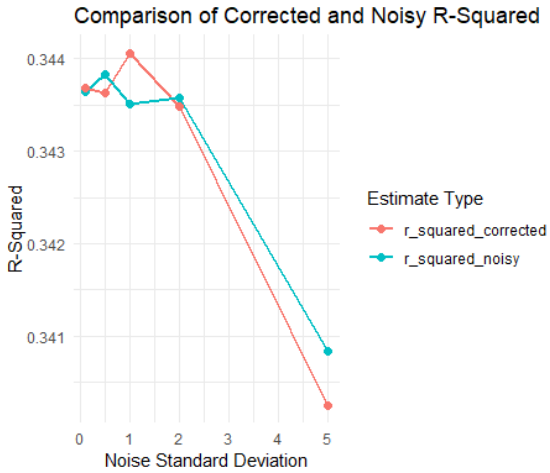

Figure 2: R-Squared : Noisy vs. Corrected

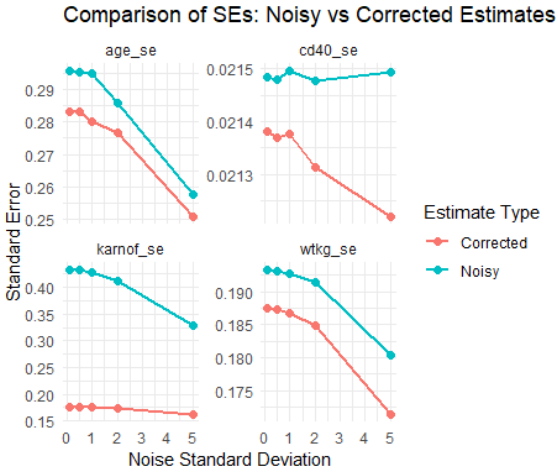

Figure 3: Standard Errors: Noisy vs. Corrected

Conclusion

This study examined the impact of noise on regression estimates, comparing noisy, corrected, and true values. Corrected estimates showed improved stability and reduced standard errors, especially for the cd40 and wtkg coefficients. However, at higher noise levels, the R-squared of corrected models declined slightly (0.34025 at noise level 5). Interestingly, noisy estimates were closer to true values at lower noise levels, highlighting a trade-off between noise correction and model fit.

References

1. Evans, G., & King, G. (2023). Statistically valid inferences from differentially private data releases, with application to the Facebook URLs dataset. Political Analysis, 31(1), 1-21.