S&DS 242/542: Theory of Statistics

Lecture 22: Inference in simple linear regression

The simple linear model

For a quantitative covariate $X \in \mathbb{R}$ and response $Y \in \mathbb{R}$, the simple linear model is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The least squares estimators are given by

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i - Y_i)^2$$

Questions:

- How to estimate standard errors for $\hat{\beta}_0, \hat{\beta}_1$?
- How to construct confidence intervals for β_0, β_1?
- How to test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$?

Inference for least squares in the normal model

The normal errors model

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are IID and distributed according to a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is normally distributed and independent of X.

Recall that

- $(\hat{\beta}_0, \hat{\beta}_1)$ are also the maximum-likelihood estimators.
- $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased estimators for (β_0, β_1). In fact, they are unbiased conditional on any fixed $x_1, \ldots, x_n$.

In the subsequent theorems, it is convenient to also condition on $x_1, \ldots, x_n$, i.e. treat these as fixed and study statistical properties of $\hat{\beta}_0, \hat{\beta}_1$ over only the randomness of $Y_1, \ldots, Y_n$.

Distribution of least-squares estimators

Theorem

Fixing $x_1, \ldots, x_n$, let $Y_1, \ldots, Y_n$ be given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\varepsilon_i \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$. Denote

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Then the least-squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ have a bivariate normal distribution

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \frac{\sigma^2}{n} \begin{pmatrix} \frac{\bar{x}^2}{s_x^2} + 1 & -\frac{\bar{x}}{s_x^2} \\ -\frac{\bar{x}}{s_x^2} & \frac{1}{s_x^2} \end{pmatrix} \right)$$

Distribution of least-squares estimators

Proof:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \left(\text{use } \begin{array}{l} Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \bar{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} \end{array}\right)$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{s_x^2}} = \beta_1 + \sum_{i=1}^{n}\underbrace{\frac{x_i - \bar{x}}{n s_x^2}}_{:= b_i} \cdot \varepsilon_i - \underbrace{\sum_{i=1}^{n}\frac{x_i - \bar{x}}{n s_x^2} \cdot \bar{\varepsilon}}_{= 0 \text{ b/c } \sum_{i=1}^{n} x_i - \bar{x} = 0}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{x} = \left(\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}\right) - \left(\beta_1 + \sum_{i=1}^{n}\frac{x_i - \bar{x}}{n s_x^2} \cdot \varepsilon_i\right)\bar{x}$$

Distribution of least-squares estimators

$$= \beta_0 + \sum_{i=1}^{n} \underbrace{\frac{1}{n}\left(1 - \frac{(x_i - \bar{x})\bar{x}}{s_x^2}\right)}_{:= a_i} \varepsilon_i$$

So: $\hat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} a_i \varepsilon_i = \beta_0 + a^T \varepsilon$, $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} b_i \varepsilon_i = \beta_1 + b^T \varepsilon$

- $(\hat{\beta}_0, \hat{\beta}_1)$ is bivariate normal b/c $\varepsilon_1, \ldots, \varepsilon_n$ are IID normal
- $\mathbb{E}\hat{\beta}_0 = \beta_0$, $\mathbb{E}\hat{\beta}_1 = \beta_1$ b/c $\mathbb{E}\varepsilon_i = 0$
- $\mathrm{Var}[\hat{\beta}_1] = \sum_{i=1}^{n} \underbrace{\mathrm{Var}[b_i \varepsilon_i]}_{= b_i^2 \cdot \sigma^2} = \sigma^2 \cdot \sum_{i=1}^{n} \frac{1}{n^2} \frac{(x_i - \bar{x})^2}{s_x^4} = \frac{\sigma^2}{n} \cdot \frac{1}{s_x^2}$

Distribution of least-squares estimators

- $$\mathrm{Var}[\hat\beta_0] = \sum_{i=1}^n \underbrace{\mathrm{Var}[a_i\varepsilon_i]}_{=a_i^2\cdot\sigma^2} = \sigma^2\cdot\sum_{i=1}^n \frac{1}{n^2}\left(1-\frac{(x_i-\bar x)\bar x}{s_x^2}\right)^2$$

$$= \sigma^2\cdot\sum_{i=1}^n \frac{1}{n^2}\left(1-\frac{2(x_i-\bar x)\bar x}{s_x^2}+\frac{(x_i-\bar x)^2\bar x^2}{s_x^4}\right)$$

$$= \sigma^2\left(\frac{1}{n}+0+\frac{\bar x^2}{n s_x^2}\right) = \frac{\sigma^2}{n}\left(1+\frac{\bar x^2}{s_x^2}\right)$$

- $$\mathrm{Cov}[\hat\beta_0,\hat\beta_1] = \sum_{i=1}^n\sum_{j=1}^n a_i b_j \underbrace{\mathrm{Cov}[\varepsilon_i,\varepsilon_j]}_{=0 \text{ unless } i=j} = \sum_{i=1}^n a_i b_i\cdot \underbrace{\mathrm{Var}[\varepsilon_i]}_{=\sigma^2}$$

$$= \sigma^2\cdot\sum_{i=1}^n \frac{1}{n}\left(1-\frac{(x_i-\bar x)\bar x}{s_x^2}\right)\cdot\frac{1}{n}\frac{x_i-\bar x}{s_x^2}$$

$$= \sigma^2\cdot\sum_{i=1}^n \frac{1}{n^2}\left(\frac{x_i-\bar x}{s_x^2}-\frac{(x_i-\bar x)^2\bar x}{s_x^4}\right) = \frac{\sigma^2}{n}\left(-\frac{\bar x}{s_x^2}\right)$$

Unbiased estimate of residual variance

Thus the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\mathrm{se}_0 = \sqrt{\frac{\sigma^2}{n}\left(\frac{\bar{x}^2}{s_x^2} + 1\right)} \quad \text{and} \quad \mathrm{se}_1 = \sqrt{\frac{\sigma^2}{n} \cdot \frac{1}{s_x^2}}$$

We can estimate these standard errors as long as we can estimate the residual variance σ^2.

The classical estimate is

$$\hat{\sigma}^2 = \frac{1}{n-2}\,\mathrm{RSS} \quad \text{where} \quad \mathrm{RSS} = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - Y_i)^2$$

Here RSS is called the **residual sum-of-squares**.

Unbiased estimate of residual variance

Theorem

Fixing $x_1, \dots, x_n$, let $Y_1, \dots, Y_n$ be given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\varepsilon_i \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$. Then $\hat{\sigma}^2$ is independent of the least-squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$, and has the (rescaled) chi-squared distribution

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2} \chi^2_{n-2}$$

Since the mean of the χ^2_{n-2} distribution is $n-2$, this implies that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$, so $\hat{\sigma}^2$ is an unbiased estimate of σ^2.

Unbiased estimate of residual variance

Proof of independence and unbiasedness:

$$\text{Let } R_i = \hat{\beta}_0 + \hat{\beta}_1 x_i - Y_i$$

$$= (\beta_0 + a^T\varepsilon) + (\beta_1 + b^T\varepsilon)x_i - (\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$= a^T\varepsilon + x_i \cdot b^T\varepsilon - \varepsilon_i \qquad (\text{Hence } \mathbb{E}R_i = 0)$$

$\Rightarrow (\hat{\beta}_0, \hat{\beta}_1, R_1, \ldots, R_n)$ is multivariate normal

- $\mathrm{Var}[\hat{\beta}_0] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{X}^2}{s_X^2}\right)$ $\qquad \mathrm{Var}[Y_i] = \mathrm{Var}[\varepsilon_i] = \sigma^2$

$\mathrm{Var}[\hat{\beta}_1] = \frac{\sigma^2}{n} \cdot \frac{1}{s_X^2}$ $\qquad \mathrm{Cov}[Y_i, \hat{\beta}_0] = \mathrm{Cov}[\varepsilon_i, a^T\varepsilon] = a_i \cdot \sigma^2$

$\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\sigma^2}{n} \cdot \frac{\bar{X}}{s_X^2}$ $\qquad \mathrm{Cov}[Y_i, \hat{\beta}_1] = \mathrm{Cov}[\varepsilon_i, b^T\varepsilon] = b_i \cdot \sigma^2$

Unbiased estimate of residual variance

- $\text{Cov}[R_i, \hat\beta_0] = \text{Cov}[\hat\beta_0 + \hat\beta_1 x_i - Y_i, \hat\beta_0]$

$$= \text{Var}[\hat\beta_0] + \text{Cov}[\hat\beta_1, \hat\beta_0]\cdot x_i - \text{Cov}[Y_i, \hat\beta_0]$$

$$= \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right) + \frac{\sigma^2}{n}\left(-\frac{\bar{x}}{s_x^2}\right)x_i - \frac{1}{n}\left(1 - \frac{(x_i - \bar{x})\bar{x}}{s_x^2}\right)\sigma^2 = 0$$

$\text{Cov}[R_i, \hat\beta_1] = \text{Cov}[\hat\beta_0 + \hat\beta_1 x_i - Y_i, \hat\beta_1]$

$$= \text{Cov}[\hat\beta_0, \hat\beta_1] + \text{Var}[\hat\beta_1]\cdot x_i - \text{Cov}[Y_i, \hat\beta_1] = 0$$

So $(R_1, \ldots, R_n)$ and $(\hat\beta_0, \hat\beta_1)$ are pairwise uncorrelated

$\Rightarrow (R_1, \ldots, R_n)$ is independent of $(\hat\beta_0, \hat\beta_1)$.

Recall $\hat\sigma^2 = \frac{1}{n-2} RSS$, $RSS = \sum_{i=1}^{n} R_i^2 \Rightarrow \hat\sigma^2$ is independent of $(\hat\beta_0, \hat\beta_1)$.

Unbiased estimate of residual variance

- $\mathbb{E}[R_i^2] = \text{Var}[R_i]$ (b/c $\mathbb{E}R_i = 0$)

$$= \text{Var}[\hat\beta_0 + \hat\beta_1 x_i - Y_i]$$

$$= \text{Var}[\hat\beta_0] + \text{Var}[\hat\beta_1]\cdot x_i^2 + \text{Var}[Y_i]$$

$$+ 2\,\text{Cov}[\hat\beta_0, \hat\beta_1]\cdot x_i - 2\,\text{Cov}[\hat\beta_0, Y_i] - 2\,\text{Cov}[\hat\beta_1, Y_i]\,x_i$$

$$= \ldots = \sigma^2 - \frac{\sigma^2}{n}\left(1 + \frac{(x_i - \bar x)^2}{s_x^2}\right)$$

$$\Rightarrow \mathbb{E}\,\text{RSS} = \sum_{i=1}^{n} \mathbb{E}[R_i^2] = n\sigma^2 - \sigma^2 - \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{(x_i - \bar x)^2}{s_x^2} = (n-2)\sigma^2$$

$\Rightarrow \mathbb{E}\,\hat\sigma^2 = \mathbb{E}\left[\frac{1}{n-2}\cdot \text{RSS}\right] = \sigma^2$ so $\hat\sigma^2$ is unbiased.

Estimated standard errors and confidence intervals

We may estimate the standard error se_1 of $\hat{\beta}_1$ by

$$\hat{\mathrm{se}}_1 = \sqrt{\frac{\hat{\sigma}^2}{n} \cdot \frac{1}{s_x^2}}$$

where $\hat{\sigma}^2 = \frac{1}{n-2}\,\mathrm{RSS}$ is the preceding estimate of residual variance.

An asymptotic confidence interval for β_1 with coverage probability $(1-\alpha)$ is given by

$$\hat{\beta}_1 \pm z^{(\alpha/2)} \cdot \hat{\mathrm{se}}_1$$

where $z^{(\alpha/2)}$ is the upper-$\alpha/2$ point of the standard normal.

Estimated standard errors and confidence intervals for β_0 are analogous.

t-test for regression coefficients

To test

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 > 0$$

observe that under H_0, we have

$$\frac{\hat{\beta}_1}{\mathrm{se}_1} \sim \mathcal{N}(0,1), \qquad \frac{\hat{\mathrm{se}}_1^2}{\mathrm{se}_1^2} = \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{1}{n-2}\,\chi^2_{n-2}$$

and these are independent. Then by definition of the t-distribution,

$$\frac{\hat{\beta}_1}{\hat{\mathrm{se}}_1} \sim t_{n-2}$$

Here $\frac{\hat{\beta}_1}{\hat{\mathrm{se}}_1}$ is the **t-statistic** for testing $H_0 : \beta_1 = 0$, and a one-sided level-α **t-test** rejects H_0 if $\frac{\hat{\beta}_1}{\hat{\mathrm{se}}_1} > t^{(\alpha)}_{n-2}$.

A two-sided t-test may be used to test against $H_1 : \beta_1 \neq 0$.

Summary

These procedures address the questions

- How to estimate standard errors for $\hat{\beta}_0, \hat{\beta}_1$?
- How to construct confidence intervals for β_0, β_1?
- How to test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$?

assuming that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and $\varepsilon_i \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$.

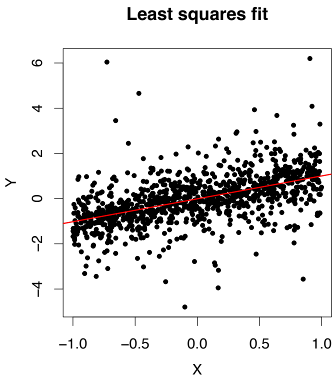
The coverage guarantee of the confidence interval and Type I error guarantee of the t-test hold for any fixed $x_1, \ldots, x_n$. Hence they also hold in a model

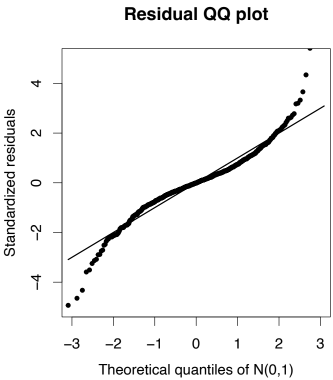
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $X_1, \ldots, X_n$ are random and $\varepsilon_1, \ldots, \varepsilon_n \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$ are independent of $X_1, \ldots, X_n$, by applying these guarantees conditional on any realization of $x_1, \ldots, x_n$.

Inference under model misspecification

Example 1: Non-normal errors





True model: $Y = X +$ heavy-tailed errors

Least-squares estimates: $(\hat{\beta}_0, \hat{\beta}_1) = (-0.009, 1.01)$
Estimated standard errors: $(\hat{\mathrm{se}}_0, \hat{\mathrm{se}}_1) = (0.030, 0.052)$

Example 1: Non-normal errors

Across 1000 simulations:

Mean of $\hat{\beta}_1$: 1.00
Standard deviation of $\hat{\beta}_1$: 0.055

Average estimated standard error $\hat{\mathrm{se}}_1$: 0.055
Empirical coverage of 90% confidence interval $\hat{\beta}_1 \pm z^{(0.05)}\,\hat{\mathrm{se}}_1$: 90%

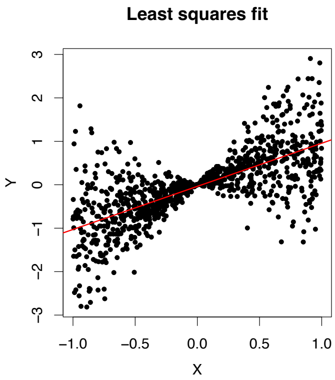
Under hypothesis $H_0 : \beta_1 = 0$ with data generated as

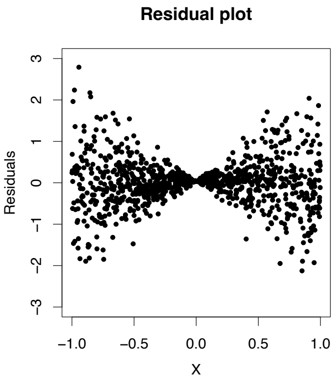
$$Y = \text{heavy-tailed errors (with no dependence on } X\text{)}$$

Empirical Type I error probability of a level-0.10 t-test: 9.9%

The preceding procedures are robust to the non-normality of errors.

Example 2: Heteroscedastic errors





True model: $Y = X + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, x^2)$ given $X = x$

Least-squares estimates: $(\hat{\beta}_0, \hat{\beta}_1) = (-0.037, 0.99)$
Estimated standard errors: $(\hat{\text{se}}_0, \hat{\text{se}}_1) = (0.018, 0.032)$

Example 2: Heteroscedastic errors

Across 1000 simulations:

Mean of $\hat{\beta}_1$: 1.00
Standard deviation of $\hat{\beta}_1$: 0.043

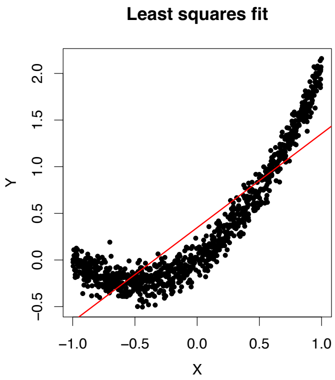
Average estimated standard error $\hat{\mathrm{se}}_1$: 0.032
Empirical coverage of 90% confidence interval $\hat{\beta}_1 \pm z^{(0.05)}\,\hat{\mathrm{se}}_1$: 76%

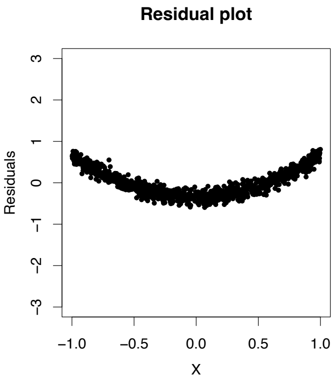
Under $H_0 : \beta_1 = 0$ with data generated as

$$Y = \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, x^2) \text{ given } X = x$$

Empirical Type I error probability of a level-0.10 t-test: 23%

Example 3: Non-linear model





True model: $Y = X + X^2 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.1)$ independent of X

Least-squares estimates: $(\hat{\beta}_0, \hat{\beta}_1) = (0.34, 1.01)$
Estimated standard errors: $(\hat{\mathrm{se}}_0, \hat{\mathrm{se}}_1) = (0.010, 0.017)$

Example 3: Non-linear model

Here $(\hat{\beta}_0, \hat{\beta}_1)$ may be understood as estimating the population least-squares coefficients

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\arg\min}\, \mathbb{E}[(\beta_0 + \beta_1 X - Y)^2] = (\tfrac{1}{3}, 1)$$

Across 1000 simulations:
Mean of $\hat{\beta}_1$: 1.00, Standard deviation of $\hat{\beta}_1$: 0.021

Average estimated standard error $\hat{\mathrm{se}}_1$: 0.017
Empirical coverage of 90% confidence interval $\hat{\beta}_1 \pm z^{(0.05)}\, \hat{\mathrm{se}}_1$: 83%

Under $H_0 : \beta_1^* = 0$ with data generated as

$$Y = X^2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.1) \text{ independent of } X$$

Empirical Type I error probability of a level-0.10 t-test: 17%

Summary

For large sample sizes n, the preceding inference procedures based on a true linear model with normal errors are reasonably robust to non-normality of the errors.

However, they are not robust to heteroscedasticity of the error variance or nonlinearity of the true regression function.

Nonparametric bootstrap estimate of standard error

There are both analytic and simulation-based approaches to correct these problems under model specification.

One approach is using the nonparametric bootstrap:

1. For each of B bootstrap simulations:
   - Resample n pairs $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ *with replacement* from $(X_1, Y_1), \ldots, (X_n, Y_n)$.
   - Compute the least squares estimators $(\hat{\beta}_0^*, \hat{\beta}_1^*)$ on the resampled data $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
2. Estimate the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ by the empirical standard deviations of $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ across the B bootstrap simulations.
3. Use these estimates instead of the model-based estimates $\hat{\mathrm{se}}_0, \hat{\mathrm{se}}_1$ to construct confidence intervals and t-statistics

Example 2: Heteroscedastic errors

Across 1000 simulations:

Mean of $\hat{\beta}_1$: 1.00
Standard deviation of $\hat{\beta}_1$: 0.043

Average bootstrap estimate of standard error: 0.042
Empirical coverage of 90% bootstrap confidence interval: 89%

Under $H_0 : \beta_1 = 0$ with data generated as

$$Y = \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, x^2) \text{ given } X = x$$

Type I error probability of level-0.10 bootstrap t-test: 9.8%

Example 3: Non-linear model

Across 1000 simulations:

Mean of $\hat{\beta}_1$: 1.00
Standard deviation of $\hat{\beta}_1$: 0.021

Average bootstrap estimated of standard error: 0.021
Empirical coverage of 90% bootstrap confidence interval: 89%

Under $H_0 : \beta_1^* = 0$ with data generated as

$$Y = X^2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.1) \text{ independent of } X$$

Type I error probability of level-0.10 bootstrap t-test: 10.9%