Simple Linear Regression

This study guide covers Simple Linear Regression. Many of the formulas are primarily derived from *Linear Regression Analysis, Fifth Edition* by Montgomery et al. and/or the materials from the STAT 6021 course taught by Dr. Woo at the University of Virginia. It is believed that the materials are well-known equations and concepts in the public domain. If you believe otherwise, please reach out to me through my Github account so that I can correct the material.

The model

For a series of observations $(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)$, we assume that the observations can be modeled by

$$y = \hat{\beta}_1 x + \hat{\beta}_0 + e$$

for appropriate $\hat{\beta}_0$ and $\hat{\beta}_1$ and an error term e. We assume assume e to be normally distributed and that $\mathrm{Var}(e) = \sigma^2$ for some fixed value of σ and $E(e) = 0$.

Basic formulas

We write the means as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

We define the fitted values for each i as

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

We define the residuals for each i as the difference between the observed value and our fitted value:

$$e_i = y_i - \hat{y}_i.$$

Calculating the slope and intercept

Let

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n}\{(y_i - \bar{y})(x_i - \bar{x})\}$$

Then

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Parameter variance

Theoretical

The variances are as below.

$$Var(\hat{y}) = \sigma^2$$

$$Var(\hat{\beta}_1) = \sigma^2/S_{xx}$$

$$Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

Estimated

We can estimate the variance of σ^2 as

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$$

We can therefore compute the estimated standard error:

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = \sqrt{\frac{SS_{Res}}{(n-2)S_{xx}}}$$

For $\hat{\beta}_0$:

$$se(\hat{\beta}_0) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

We can similarly estimate the other standard errors.

Sums of squares

The total sum of squares is

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The regression sum of squares is

$$SS_R = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$$

The residual sum of squares is

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

We have

$$SS_T = SS_R + SS_{Res}$$

. Also

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$$

Distributions

If we assume $H_0 : \beta_1 = 0$, then we have an $F_{1,n-2}$ distribution with the following statistic:

$$F_0 = MS_R/MS_{Res}$$

If we assume $\beta_1 \neq 0$ then we have a non-central $F_{1,n-2}$ distribution with a non-centrality parameter of

$$\lambda = \beta_1^2 S_{xx}/\sigma^2$$

These two together justify the F test.

Other distributions justify additional the tests and confidence interval calculations below.

Hypothesis tests

When using a two-sided hypothesis test: $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. The null hypothesis is that our observations could be explained despite a slope of zero, indicating a lack of (linear) relationship between the x and y values. The alternative hypothesis is that there is a non-zero slope.

We can the t statistic For β_1:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}}$$

where β_{10} is the hypothetical value and $\hat{\beta}_0$ is the value obtained for the model. That is, for the test $H_0 : \beta_0 = 0$, we set $\beta_{10} = 0$ and reject the null hypothesis if $|t_0| > t_{a/2,n-2}$ where n is the number of observations.

We also reproduce the t statistic for β_0:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}}$$

Confidence Interval Formulas

Note: We use the estimates for the standard deviation here. If you actually know the population σ, you shouldn't use these formulas.

Confidence interval for β_1:

$$\hat{\beta}_1 - t_{a/2,n-2}\sqrt{\frac{SS_{Res}}{(n-2)S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{a/2,n-2}\sqrt{\frac{SS_{Res}}{(n-2)S_{xx}}}$$

Confidence interval for β_0:

$$\hat{\beta}_0 - t_{a/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \leq \beta_0 \leq \hat{\beta}_0 + t_{a/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Confidence interval for σ^2:

$$\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}$$

Confidence interval for expected value $E(y|x_0)$:

$$\hat{\mu}_{y|x_0} - t_{a/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq E(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{a/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

where $\hat{\mu}_{y|x0}$, also written as $\widehat{E(y|x_0)}$, is estimated by $\hat{y}_{x=x_0} = \beta_1 x_0 + \beta_0$ (eq. 2.42 in the book).

Confidence interval for predicted value y_0:

$$\hat{y}_0 - t_{a/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{a/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Confidence interval for the average of m-many predicted values $(y_0)_1, (y_0)_2, ..., (y_0)_m$:

$$\hat{y}_0 - t_{a/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{a/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Definitions

- NID(μ, σ^2) is shorthand for "normally and independently distributed with mean μ and variance σ^2.
- BLUE means "best linear unbiased estimators"
- SLR means "simple linear regression"
- Model adequacy

Special formulas

- $\sum_{i=1}^{n} e_i = 0$ (that is, the residuals cancel out)
- $\sum_{i=1}^{n} x_i e_i = 0$
- $\sum_{i=1}^{n} y_i e_i = 0$
- $Cov(\bar{y}, \hat{\beta}_1) = 0$
- Approximate $E(R^2)$:

$$E(R^2) \approx \frac{\beta_1^2 S_{xx}/n - 1}{\frac{\beta_1^2 S_{xx}}{n-1} + \sigma^2}$$