Simple Linear Regression

Emilija Perković

Dept. of Statistics
University of Washington

Simple linear regression in Weisberg's notation

Mean Function:

$$\mathrm{E}[Y|X=x] = \beta_0 + \beta_1 x,$$

Variance Function:

$$\mathrm{Var}[Y|X=x] = \sigma^2,$$

where

Simple linear regression in Weisberg's notation

Mean Function:

$$E[Y|X=x] = \beta_0 + \beta_1 x,$$

Variance Function:

$$\mathrm{Var}[Y|X=x] = \sigma^2,$$

where

- β_0 is the intercept,

Simple linear regression in Weisberg's notation

Mean Function:

$$E[Y|X = x] = \beta_0 + \beta_1 x,$$

Variance Function:

$$\mathrm{Var}[Y|X = x] = \sigma^2,$$

where

- β_0 is the intercept,
- β_1 is the slope, and

Simple linear regression in Weisberg's notation

Mean Function:

$$E[Y|X=x] = \beta_0 + \beta_1 x,$$

Variance Function:

$$\mathrm{Var}[Y|X=x] = \sigma^2,$$

where

- β_0 is the intercept,
- β_1 is the slope, and
- $0 < \sigma^2 < \infty$ is the variance of Y .

Simple linear regression in Weisberg's notation

Mean Function:

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x,$$

Variance Function:

$$\mathrm{Var}[Y|X = x] = \sigma^2,$$

where

- β_0 is the intercept,
- β_1 is the slope, and
- $0 < \sigma^2 < \infty$ is the variance of Y .

Other common notation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } i = 1, \ldots, n, \epsilon_i \text{ iid, with}$$
$$\mathrm{E}[\epsilon_i|X = x] = 0 \text{ and } \mathrm{Var}[\epsilon_i|X = x] = \sigma^2.$$

Simple linear regression: Assumptions

$$y = f(x) + \epsilon$$

Regression Assumptions:

Simple linear regression: Assumptions

$$y = f(x) + \epsilon$$

Regression Assumptions:

- Variance of Y does not depend on X (homoscedasticity).
- Errors $\epsilon = y - \mathrm{E}[Y|X = x]$ have zero mean, i.e., $\mathrm{E}[\epsilon|X = x] = 0$.

Simple linear regression: Assumptions

$$y = f(x) + \epsilon$$

Regression Assumptions:

- Variance of Y does not depend on X (homoscedasticity).
- Errors $\epsilon = y - \mathrm{E}[Y|X = x]$ have zero mean, i.e., $\mathrm{E}[\epsilon|X = x] = 0$.
- Errors ϵ are independent (the error for one case gives no information about the error for another case).

Simple linear regression: Assumptions

$$y = f(x) + \epsilon$$

Regression Assumptions:

- Variance of Y does not depend on X (homoscedasticity).
- Errors $\epsilon = y - \mathrm{E}[Y|X = x]$ have zero mean, i.e., $\mathrm{E}[\epsilon|X = x] = 0$.
- Errors ϵ are independent (the error for one case gives no information about the error for another case).
- Errors are assumed to be normally distributed.
  Note: The normality assumption is much stronger than we need in many cases (e.g., see Weisberg p.22). It is used primarily for inference (tests and confidence intervals) with small sample sizes.

OLS Estimation

Given a set of data points $(x_1, y_1), \ldots, (x_n, y_n)$, we learn about β_0 and β_1 by obtaining estimates of β_0 and β_1 from the data.

OLS Estimation

Given a set of data points $(x_1, y_1), \ldots, (x_n, y_n)$, we learn about β_0 and β_1 by obtaining estimates of β_0 and β_1 from the data.

One way to estimate β_0 and β_1 is to find values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the residual sum of squares:

$$
\begin{aligned}
RSS(\beta_0, \beta_1) &= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \\
&= \sum_{i=1}^{n} \epsilon_i^2
\end{aligned}
$$

OLS estimation

One can obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ by

OLS estimation

One can obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ by

- setting the partial derivatives of RSS (residual sum of squares) with respect to β_0 and β_1 equal to zero

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{n}(2\beta_0 - 2y_i + 2\beta_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = \sum_{i=1}^{n}(2\beta_1 x_i^2 - 2y_i + 2\beta_0 x_i) = 0$$

OLS estimation

One can obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ by

- setting the partial derivatives of RSS (residual sum of squares) with respect to β_0 and β_1 equal to zero

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{n}(2\beta_0 - 2y_i + 2\beta_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = \sum_{i=1}^{n}(2\beta_1 x_i^2 - 2y_i + 2\beta_0 x_i) = 0$$

- and **solving** these normal equations.

OLS estimation

One can obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ by

- setting the partial derivatives of RSS (residual sum of squares) with respect to β_0 and β_1 equal to zero

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{n} (2\beta_0 - 2y_i + 2\beta_1 x_i) = 0$$
$$\frac{\partial RSS}{\partial \beta_1} = \sum_{i=1}^{n} (2\beta_1 x_i^2 - 2y_i + 2\beta_0 x_i) = 0$$

- and **solving** these normal equations.

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained in such a way are called **ordinary least squares estimates** (OLS estimates) of β_0 and β_1.

The 'hat' operator

Question: What is the conceptual difference between β_0 and $\hat{\beta}_0$?

The ‘hat’ operator

Question: What is the conceptual difference between β_0 and $\hat{\beta}_0$?

We use the hat operator to **distinguish between parameters and their estimates**.

The 'hat' operator

Question: What is the conceptual difference between β_0 and $\hat{\beta}_0$?

We use the hat operator to **distinguish between parameters and their estimates**.

For example:

- Errors: $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$, $i = 1, \dots, n$,
- Residuals: $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, $i = 1, \dots, n$.

The 'hat' operator

Question: What is the conceptual difference between β_0 and $\hat{\beta}_0$?

We use the hat operator to **distinguish between parameters and their estimates**.

For example:

- Errors: $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$, $i = 1, \ldots, n$,
- Residuals: $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, $i = 1, \ldots, n$.

And also:

- Observed value: y_i, $i = 1, \ldots, n$,
- Fitted value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$.

The OLS estimates for slope and intercept

$$\hat{\beta}_1 = \frac{SXY}{SXX}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

where SXY is the sum of cross-products of the deviations of x_i and y_i from their means:

$$SXY = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

and SXX is the sum of squared deviations of x_i from the sample mean of x:

$$SXX = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

The OLS estimates for slope and intercept

Note that since the sampling variance of X is

$$SD_x^2 = \frac{SXX}{n-1}.$$

The OLS estimates for slope and intercept

Note that since the sampling variance of X is

$$SD_x^2 = \frac{SXX}{n-1}.$$

And the sampling covariance is:

$$s_{xy} = \frac{SXY}{n-1}.$$

The OLS estimates for slope and intercept

Note that since the sampling variance of X is

$$SD_x^2 = \frac{SXX}{n-1}.$$

And the sampling covariance is:

$$s_{xy} = \frac{SXY}{n-1}.$$

Then

$$\hat{\beta}_1 = \frac{s_{xy}}{SD_x^2} = \frac{SXY(n-1)}{(n-1)SXX} = \frac{SXY}{SXX}.$$

The OLS estimates for slope and intercept

Note that the OLS regression line goes through the point $(\overline{x}, \overline{y})$, the center mass of the data.

The OLS estimates for slope and intercept

Note that the OLS regression line goes through the point $(\overline{x}, \overline{y})$, the center mass of the data.

Verify by plugging in $\overline{x}$, $\overline{y}$ and the OLS estimates into the mean function for the simple regression:

$$\overline{y} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x}.$$

The OLS estimates for slope and intercept

Note that the OLS regression line goes through the point $(\overline{x}, \overline{y})$, the center mass of the data.

Verify by plugging in $\overline{x}$, $\overline{y}$ and the OLS estimates into the mean function for the simple regression:

$$\overline{y} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x}.$$

Interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$: Fitting the regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2), \epsilon_i \text{ iid},$$

we obtain estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.

The OLS estimates for slope and intercept

Note that the OLS regression line goes through the point $(\overline{x}, \overline{y})$, the center mass of the data.
Verify by plugging in $\overline{x}$, $\overline{y}$ and the OLS estimates into the mean function for the simple regression:

$$\overline{y} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x}.$$

Interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$: Fitting the regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2), \epsilon_i \text{ iid},$$

we obtain estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.

- $\hat{\beta}_0$ - The estimated average value of y for $x = 0$,
- $\hat{\beta}_1$ - For every unit increase of x, we estimate that y increases by $\hat{\beta}_1$ on average.

Example: Forbes data

Find the OLS estimates for the regression of pressure on temperature, given:

$$\bar{x} = 202.9529$$

$$\bar{y} = 25.05882$$

$$SXX = 530.7824$$

$$SXY = 277.5421$$

Example: Forbes data

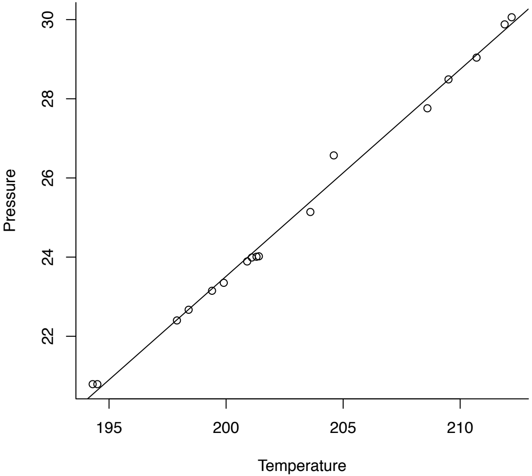
Find the OLS estimates for the regression of pressure on temperature, given:

$$\bar{x} = 202.9529$$
$$\bar{y} = 25.05882$$
$$SXX = 530.7824$$
$$SXY = 277.5421$$

Using the formulae for OLS estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$\hat{\beta}_1 = \frac{277.5421}{530.7824} \approx 0.523$$
$$\hat{\beta}_0 = 25.05882 - 0.523 * 202.9529 \approx -81.064$$

Example: Forbes data



OLS properties

Property 1.

$\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as linear functions of $y_1, \ldots, y_n$, e.g.,

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i, \text{ where } c_i = \frac{x_i - \bar{x}}{SXX}.$$

OLS properties

Property 1.

$\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as linear functions of $y_1, \ldots, y_n$, e.g.,

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i, \text{ where } c_i = \frac{x_i - \overline{x}}{SXX}.$$

Proof:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{SXX} \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{SXX} - \overline{y}\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{SXX} \\
&= \sum_{i=1}^{n} c_i y_i - \frac{\overline{y}}{SXX}\left(\frac{n\sum_{i=1}^{n} x_i}{n} - n\overline{x}\right) = \sum_{i=1}^{n} c_i y_i
\end{aligned}$$

OLS properties

For OLS estimate of the intercept $\hat{\beta}_0$ recall

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

OLS properties

For OLS estimate of the intercept $\hat{\beta}_0$ recall

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

Since the sample mean of y, $\overline{y}$, is a linear combination of $y_1, \ldots, y_n$, and we just showed that $\hat{\beta}_1$ is a linear combination of $y_1, \ldots, y_n$, then $\hat{\beta}_0$ is a linear combination of $y_1, \ldots, y_n$ as well.

OLS properties

For OLS estimate of the intercept $\hat{\beta}_0$ recall

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

Since the sample mean of y, $\overline{y}$, is a linear combination of $y_1, \ldots, y_n$, and we just showed that $\hat{\beta}_1$ is a linear combination of $y_1, \ldots, y_n$, then $\hat{\beta}_0$ is a linear combination of $y_1, \ldots, y_n$ as well.

Exercise: Find d_i, $i = 1, \ldots, n$ such that $\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i$.

OLS properties

Property 2. If $\mathrm{E}[\epsilon_i | X = x] = 0$, for all $i = 1, \ldots, n$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, that is, $\mathrm{E}[\hat{\beta}_0 | X = x] = \beta_0$ and $\mathrm{E}[\hat{\beta}_1 | X = x] = \beta_1$.

OLS properties

Property 2. If $\mathrm{E}[\epsilon_i|X=x]=0$, for all $i=1,\dots,n$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, that is, $\mathrm{E}[\hat{\beta}_0|X=x]=\beta_0$ and $\mathrm{E}[\hat{\beta}_1|X=x]=\beta_1$.

Proof:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_1|X=x] &= \mathrm{E}[\sum_{i=1}^{n} c_i y_i|X=x] = \sum_{i=1}^{n} c_i \mathrm{E}[y_i|X=x] \\
&= \sum_{i=1}^{n} c_i(\beta_0+\beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\
&= \frac{\beta_0}{SXX}\sum_{i=1}^{n}(x_i-\overline{x}) + \beta_1 \sum_{i=1}^{n} \frac{x_i(x_i-\overline{x})}{SXX}
\end{aligned}$$

OLS properties

Proof continued: (For $\hat{\beta}_1$, given $X = x$)

$$
\begin{aligned}
\mathrm{E}[\hat{\beta}_1 | X = x] &= \frac{\beta_0}{SXX} \sum_{i=1}^{n} (x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n} \frac{x_i (x_i - \overline{x})}{SXX} \\
&= \beta_1 \frac{\sum_{i=1}^{n} x_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})} \\
&= \beta_1 \frac{[\sum_{i=1}^{n} x_i (x_i - \overline{x}) - \overline{x}(x_i - \overline{x}) + \overline{x}(x_i - \overline{x})]}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})} \\
&= \beta_1 \frac{[\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})]}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})} + \beta_1 \frac{\sum_{i=1}^{n} \overline{x}(x_i - \overline{x})}{SXX} \\
&= \beta_1
\end{aligned}
$$

For the last step, we use the same trick as before to show that the second term is 0.

OLS properties

Exercise: Show that $\hat{\beta}_0$ is an unbiased estimator of β_0.

OLS properties

Exercise: Show that $\hat{\beta}_0$ is an unbiased estimator of β_0.

Property 3. If $\mathrm{E}[\epsilon_i|X=x]=0$, $\mathrm{Var}[\epsilon_i|X=x]=\sigma^2$ and the errors ϵ_i are uncorrelated for all $i=1,\ldots,n$,
then the variances of the OLS estimators are:

$$\mathrm{Var}[\hat{\beta}_1|X=x]=\frac{\sigma^2}{SXX},$$

$$\mathrm{Var}[\hat{\beta}_0|X=x]=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SXX}\right),$$

$$\mathrm{Cov}[\hat{\beta}_0,\hat{\beta}_1|X=x]=-\sigma^2\frac{\overline{x}}{SXX}.$$

OLS properties

Exercise: Show that $\hat{\beta}_0$ is an unbiased estimator of β_0.

Property 3. If $\mathrm{E}[\epsilon_i|X=x]=0$, $\mathrm{Var}[\epsilon_i|X=x]=\sigma^2$ and the errors ϵ_i are uncorrelated for all $i=1,\dots,n$,
then the variances of the OLS estimators are:

$$\mathrm{Var}[\hat{\beta}_1|X=x]=\frac{\sigma^2}{SXX},$$

$$\mathrm{Var}[\hat{\beta}_0|X=x]=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SXX}\right),$$

$$\mathrm{Cov}[\hat{\beta}_0,\hat{\beta}_1|X=x]=-\sigma^2\frac{\overline{x}}{SXX}.$$

Proof: Exercise.

OLS properties

Property 4.

The sum of residuals form an OLS fit is zero (as long as $\beta_0 \neq 0$):

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0.$$

Proof: Exercise.

OLS properties

Gauss-Markov Theorem. Assume $\mathrm{E}[\epsilon_i|X = x] = 0$, $\mathrm{Var}[\epsilon_i|X = x] = \sigma^2$ and the errors ϵ_i are uncorrelated for all $i = 1, \ldots, n$.
Among all unbiased estimators that are linear combinations of y's, the OLS estimators of regression coefficients have the smallest variance, i.e., they are **b**est **l**inear **u**nbiased **e**stimators (BLUE).

OLS properties

Gauss-Markov Theorem. Assume $\mathrm{E}[\epsilon_i|X = x] = 0$, $\mathrm{Var}[\epsilon_i|X = x] = \sigma^2$ and the errors ϵ_i are uncorrelated for all $i = 1, \ldots, n$.
Among all unbiased estimators that are linear combinations of y's, the OLS estimators of regression coefficients have the smallest variance, i.e., they are **b**est **l**inear **u**nbiased **e**stimators (BLUE).

- Note 1: The Gauss-Markov Theorem as stated does not require the assumption of normality of the error terms.

OLS properties

Gauss-Markov Theorem. Assume $\mathrm{E}[\epsilon_i|X=x]=0$, $\mathrm{Var}[\epsilon_i|X=x]=\sigma^2$ and the errors ϵ_i are uncorrelated for all $i=1,\ldots,n$.
Among all unbiased estimators that are linear combinations of y's, the OLS estimators of regression coefficients have the smallest variance, i.e., they are **b**est **l**inear **u**nbiased **e**stimators (BLUE).

- Note 1: The Gauss-Markov Theorem as stated does not require the assumption of normality of the error terms. Adding the assumption of normality of the errors, one can show that OLS estimators are BLUE estimators among all unbiased estimators (not only linear functions of y's).

OLS properties

Gauss-Markov Theorem. Assume $\mathrm{E}[\epsilon_i|X=x]=0$, $\mathrm{Var}[\epsilon_i|X=x]=\sigma^2$ and the errors ϵ_i are uncorrelated for all $i=1,\ldots,n$.
Among all unbiased estimators that are linear combinations of y's, the OLS estimators of regression coefficients have the smallest variance, i.e., they are **b**est **l**inear **u**nbiased **e**stimators (BLUE).

- Note 1: The Gauss-Markov Theorem as stated does not require the assumption of normality of the error terms. Adding the assumption of normality of the errors, one can show that OLS estimators are BLUE estimators among all unbiased estimators (not only linear functions of y's).
- Note 2: The Gauss-Markov Theorem does not tell one to use least squares all the time, but it strongly suggests it.

Residuals

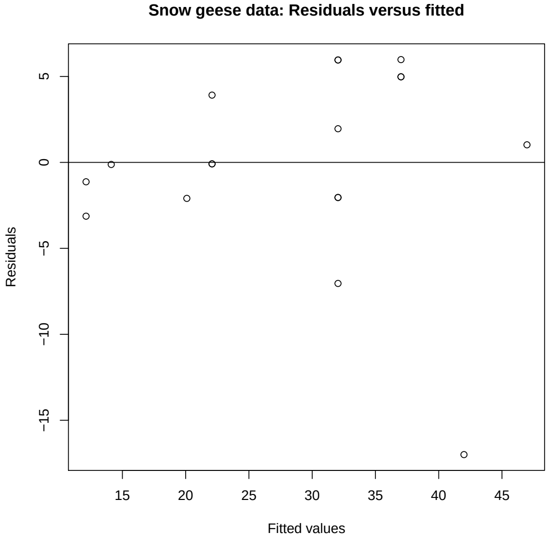
$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Let's examine plot of residuals versus fitted values for the Snow Geese data for violations of the regression assumptions.

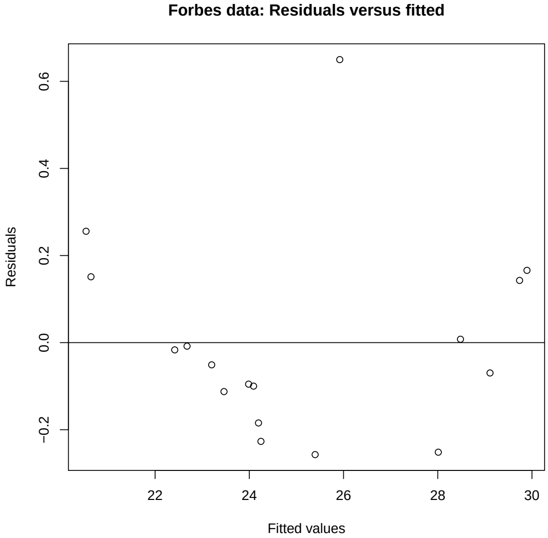
Things we are looking for:

- Curvature of the mean trend (indicates that the mean function is inappropriate);
- Increase or decrease in magnitude when fitted values are increasing (indicates non-constant variance);
- Residuals that are large in magnitude compared to the rest (indicates outliers).

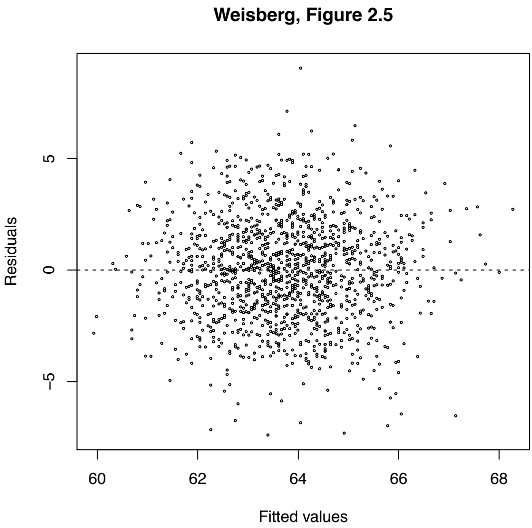
Snow geese data: Residual plot



Forbes data: Residual plot



Heights data: Residual plot



Residual assumption violations

Outliers: What to do with outliers?

In some cases we may know about specific reasons why an outlier was observed. You should not simply remove an outlier from your data without careful consideration.

Residual assumption violations

Outliers: What to do with outliers?
In some cases we may know about specific reasons why an outlier was observed. You should not simply remove an outlier from your data without careful consideration.

Mean trend and non-constant variance:
We will address some remedies for dealing with curvature in the mean trend and with non-constant variance later in the class.

Normality:
To check for the normality of the errors you can use histograms or normal qq-plots, these will be discussed later in the course.

Independence:
Plot residuals versus index and look for trends. Alternatives: turning point test, runs test, portmanteau test, Durbin-Watson test etc.

Estimating the Residual Variance

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on σ^2 (see e.g., Property 3.).
However, in many cases σ^2 is unknown.

Estimating the Residual Variance

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on σ^2 (see e.g., Property 3.).

However, in many cases σ^2 is unknown. Solution: Estimate σ^2.

Estimating the Residual Variance

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on σ^2 (see e.g., Property 3.). However, in many cases σ^2 is unknown. Solution: Estimate σ^2.

Assuming the errors are uncorrelated and have zero mean and common variance σ^2, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{d.f.},$$

where $d.f.$ stands for *degrees of freedom*.

Estimating the Residual Variance

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on σ^2 (see e.g., Property 3.). However, in many cases σ^2 is unknown. Solution: Estimate σ^2.

Assuming the errors are uncorrelated and have zero mean and common variance σ^2, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{d.f.},$$

where $d.f.$ stands for *degrees of freedom*.

residual d.f. = ([number of samples] - [number of parameters we are estimating])

Estimating the Residual Variance

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on σ^2 (see e.g., Property 3.). However, in many cases σ^2 is unknown. Solution: Estimate σ^2.

Assuming the errors are uncorrelated and have zero mean and common variance σ^2, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{d.f.},$$

where $d.f.$ stands for *degrees of freedom*.

residual d.f. = ([number of samples] - [number of parameters we are estimating])

Why?

Estimating the Residual Variance

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on σ^2 (see e.g., Property 3.). However, in many cases σ^2 is unknown. Solution: Estimate σ^2.

Assuming the errors are uncorrelated and have zero mean and common variance σ^2, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{d.f.},$$

where $d.f.$ stands for *degrees of freedom*.

residual d.f. = ([number of samples] - [number of parameters we are estimating])

Why? Because estimating parameters imposes constraints, e.g.,

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{n} (2\beta_0 - 2y_i + 2\beta_1 x_i) = 0$$

Estimating the Residual Variance

How many residual degrees of freedom does a simple regression with n samples have?

Estimating the Residual Variance

How many residual degrees of freedom does a simple regression with n samples have?
For simple linear regression with n samples, the number of residual degrees of freedom is $n-2$.

Estimating the Residual Variance

How many residual degrees of freedom does a simple regression with n samples have?

For simple linear regression with n samples, the number of residual degrees of freedom is $n-2$.

Our estimate of the residual variance for simple regression is then:

$$\hat{\sigma}^2 = \frac{RSS}{n-2},$$

Estimating the Residual Variance

How many residual degrees of freedom does a simple regression with n samples have?

For simple linear regression with n samples, the number of residual degrees of freedom is $n-2$.

Our estimate of the residual variance for simple regression is then:

$$\hat{\sigma}^2 = \frac{RSS}{n-2},$$

If $\mathrm{E}[\epsilon_i|X=x]=0$, $\mathrm{Var}[\epsilon_i|X=x]=\sigma^2$ and the errors ϵ_i are uncorrelated for all $i=1,\ldots,n$, then

Estimating the Residual Variance

How many residual degrees of freedom does a simple regression with n samples have?

For simple linear regression with n samples, the number of residual degrees of freedom is $n-2$.

Our estimate of the residual variance for simple regression is then:

$$\hat{\sigma}^2 = \frac{RSS}{n-2},$$

If $\mathrm{E}[\epsilon_i | X = x] = 0$, $\mathrm{Var}[\epsilon_i | X = x] = \sigma^2$ and the errors ϵ_i are uncorrelated for all $i = 1, \ldots, n$, then

$$\mathrm{E}[\hat{\sigma}^2 | X = x] = \sigma^2,$$

the estimate is unbiased.

Estimating the Residual Variance

Note also that

$$RSS = RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = SYY - \hat{\beta}_1^2 SXX = SYY - \frac{SXY^2}{SXX},$$

Estimating the Residual Variance

Note also that

$$RSS = RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = SYY - \hat{\beta}_1^2 SXX = SYY - \frac{SXY^2}{SXX},$$

where

- $SYY = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the *total sum of squares* (total amount of variability in the response) and

Estimating the Residual Variance

Note also that

$$RSS = RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = SYY - \hat{\beta}_1^2 SXX = SYY - \frac{SXY^2}{SXX},$$

where

- $SYY = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the *total sum of squares* (total amount of variability in the response) and
- $\frac{SXY^2}{SXX}$ is the regression sum of squares (the difference between the total and the residual sums of squares).

Estimating the Residual Variance

Note also that

$$RSS = RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = SYY - \hat{\beta}_1^2 SXX = SYY - \frac{SXY^2}{SXX},$$

where

- $SYY = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the *total sum of squares* (total amount of variability in the response) and
- $\frac{SXY^2}{SXX}$ is the regression sum of squares (the difference between the total and the residual sums of squares).

Then for the Forbes' data with

$$RSS = 0.813143, \; n = 17,$$

the estimated residual variance is

Estimating the Residual Variance

Note also that

$$RSS = RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = SYY - \hat{\beta}_1^2 SXX = SYY - \frac{SXY^2}{SXX},$$

where

- $SYY = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the *total sum of squares* (total amount of variability in the response) and
- $\frac{SXY^2}{SXX}$ is the regression sum of squares (the difference between the total and the residual sums of squares).

Then for the Forbes' data with

$$RSS = 0.813143, \; n = 17,$$

the estimated residual variance is

$$\hat{\sigma}^2 = \frac{0.813143}{17 - 2} \approx 0.054.$$

Standard Errors of the OLS Estimators

The square root of an estimated variance is called *standard error*.

Standard Errors of the OLS Estimators

The square root of an estimated variance is called *standard error*.

Since the true value of σ^2 is unknown, we replace σ^2 with its unbiased estimate, $\hat{\sigma}^2$, to obtain standard errors of the regression coefficients:

Standard Errors of the OLS Estimators

The square root of an estimated variance is called *standard error*.

Since the true value of σ^2 is unknown, we replace σ^2 with its unbiased estimate, $\hat{\sigma}^2$, to obtain standard errors of the regression coefficients:

$$SE(\hat{\beta}_1|X=x) = \frac{\hat{\sigma}}{\sqrt{SXX}}$$

$$SE(\hat{\beta}_0|X=x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}}.$$

Standard Errors of the OLS Estimators

The square root of an estimated variance is called *standard error*.

Since the true value of σ^2 is unknown, we replace σ^2 with its unbiased estimate, $\hat{\sigma}^2$, to obtain standard errors of the regression coefficients:

$$SE(\hat{\beta}_1|X=x) = \frac{\hat{\sigma}}{\sqrt{SXX}}$$

$$SE(\hat{\beta}_0|X=x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}}.$$

Distribution of estimates

Let us come back to simple linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Distribution of estimates

Let us come back to simple linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Then,

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

Distribution of estimates

Let us come back to simple linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Then,

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of $y_1, \ldots, y_n$ (Property 1), then $(\hat{\beta}_0, \hat{\beta}_1)$ follows a **bivariate normal** distribution.

Confidence Intervals

Because $(\hat{\beta}_0, \hat{\beta}_1)$ follow a bivariate normal distribution, when σ^2 is known, the marginal distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are univariate normal.

Confidence Intervals

Because $(\hat{\beta}_0, \hat{\beta}_1)$ follow a bivariate normal distribution, when σ^2 is known, the marginal distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are univariate normal.

$$\hat{\beta}_0 | X = x \sim \mathcal{N}(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})),$$

given that $\epsilon_i | X = x$ iid $\mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$.

Confidence Intervals

Because $(\hat{\beta}_0, \hat{\beta}_1)$ follow a bivariate normal distribution, when σ^2 is known, the marginal distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are univariate normal.

$$\hat{\beta}_0 | X = x \sim \mathcal{N}(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})),$$

given that $\epsilon_i | X = x$ iid $\mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Then

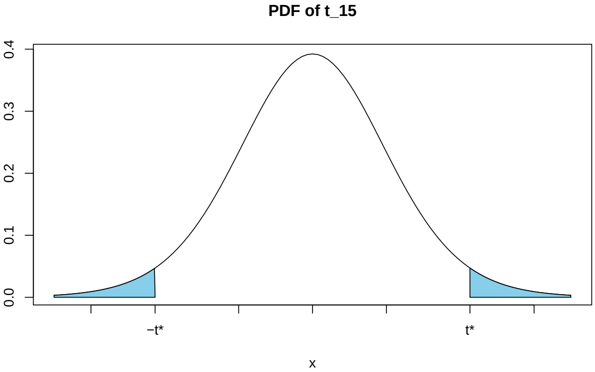
$$\frac{\hat{\beta}_0 - \beta_0}{\sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})} | X = x \sim \mathcal{N}(0, 1).$$

Since σ^2 is usually not known and is instead estimated as $\hat{\sigma}^2$,

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})} | X = x \sim t_{n-2}.$$

The t-distribution with $n - 2$ degrees of freedom is the appropriate reference distribution for constructing the confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$.

$n = 17$ and we are interested in a 90% CI for β_0



$P\left(\frac{\hat{\beta}_0-\beta_0}{\sqrt{\hat{\sigma}^2/SXX}} \leq |t^*| \middle| X=x\right) = 0.9$, so $t^* = t_{0.95,15}$. Then

$$P\left(-t^* \leq \frac{\hat{\beta}_0-\beta_0}{SE(\hat{\beta}_0|X=x)} \leq t^*\right) = 0.9.$$

Confidence Interval for $\hat{\beta}_0$

Since

$$P(-t^* \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X=x)} \leq t^*|X=x) = 0.9,$$

Confidence Interval for $\hat{\beta}_0$

Since

$$P\left(-t^* \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X=x)} \leq t^* | X = x\right) = 0.9,$$

$$P(\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x) \leq \beta_0 \leq \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)|X=x) = 0.9,$$

a 90% confidence interval for $\hat{\beta}_0$ when $n = 17$ is:

$$\left[\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x), \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)\right]$$

Confidence Interval for $\hat{\beta}_0$

Since

$$P(-t^* \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X=x)} \leq t^* | X = x) = 0.9,$$

$$P(\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x) \leq \beta_0 \leq \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x) | X = x) = 0.9,$$

a 90% confidence interval for $\hat{\beta}_0$ when $n = 17$ is:

$$\left[\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x), \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)\right]$$

The general form of a two-sided $(1-\alpha) \times 100\%$ confidence interval for a symmetric probability distribution is:

Estimate ± $(1-\alpha/2)$-quantile of the prob. dist. × SE of estimate.

Confidence Interval for $\hat{\beta}_0$

Since

$$P\left(-t^* \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X=x)} \leq t^* | X = x\right) = 0.9,$$

$$P(\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x) \leq \beta_0 \leq \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)|X=x) = 0.9,$$

a 90% confidence interval for $\hat{\beta}_0$ when $n = 17$ is:

$$\left[\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x), \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)\right]$$

The general form of a two-sided $(1-\alpha) \times 100\%$ confidence interval for a symmetric probability distribution is:

Estimate ± $(1-\alpha/2)$-quantile of the prob. dist. × SE of estimate.

The interpretation of confidence intervals is based on repeated sampling.

Confidence Interval for $\hat{\beta}_0$

Since

$$P(-t^* \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X=x)} \leq t^*|X=x) = 0.9,$$

$$P(\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x) \leq \beta_0 \leq \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)|X=x) = 0.9,$$

a 90% confidence interval for $\hat{\beta}_0$ when $n = 17$ is:

$$\left[\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X=x), \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X=x)\right]$$

The general form of a two-sided $(1-\alpha) \times 100\%$ confidence interval for a symmetric probability distribution is:

Estimate ± $(1-\alpha/2)$-quantile of the prob. dist. × SE of estimate.

The interpretation of confidence intervals is based on repeated sampling. If samples of size n are drawn repeatedly and, say, 95% confidence intervals are estimated for the intercept, then 95% of those intervals (on average) would contain the true parameter β_0.

Example: Confidence Interval for $\hat{\beta}_0$

Forbes data ($n = 17$), regression of pressure on temperature.

Example: Confidence Interval for $\hat{\beta}_0$

Forbes data ($n = 17$), regression of pressure on temperature.

Given that

$$\hat{\beta}_0 = -81.064,$$

$$SE(\hat{\beta}_0|X = x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}} = 2.052 \text{ and}$$

$$t_{0.95,15} = 1.753,$$

find the 90% confidence interval for the intercept.

Example: Confidence Interval for $\hat{\beta}_0$

Forbes data ($n = 17$), regression of pressure on temperature.

Given that

$$\hat{\beta}_0 = -81.064,$$

$$SE(\hat{\beta}_0 | X = x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}} = 2.052 \text{ and}$$

$$t_{0.95,15} = 1.753,$$

find the 90% confidence interval for the intercept.

The 90% confidence interval for the intercept is

$$-84.661 \leq \beta_0^* \leq -77.467$$

Example: Confidence Interval for $\hat{\beta}_0$

Forbes data ($n = 17$), regression of pressure on temperature.

Given that

$$\hat{\beta}_0 = -81.064,$$

$$SE(\hat{\beta}_0 | X = x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}} = 2.052 \text{ and}$$

$$t_{0.95,15} = 1.753,$$

find the 90% confidence interval for the intercept.

The 90% confidence interval for the intercept is

$$-84.661 \leq \beta_0^* \leq -77.467$$

Interpret.