Simple Linear Regression

We will begin by considering the SIMPLE LINEAR REGRESSION MODEL. A single response variable Y is related to a single predictor variable X via the linear equation

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

β 0 & 1 are fixed, X could ether be fixed or random but usually fixed, ε is random, so Y is random

Here, the intercept β_0 and the slope β_1 are assumed to be unknown constants and the residual terms ϵ are assumed to be independent, normally distributed random variables with mean zero and unknown (but constant) variance σ^2.

The predictor variable X is usually assumed to be controllable (measured without error), while the response variable Y is a random variable whose distribution depends on the value of X.

$$E(Y|X) = \beta_0 + \beta_1 X$$

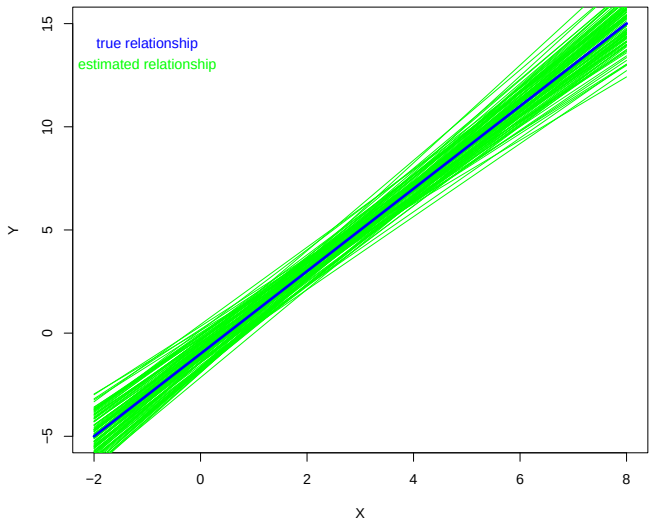
σ is the error stand deviation

and

$$Var(Y|X) = Var(\beta_0 + \beta_1 X + \epsilon) = Var(\epsilon) = \sigma^2$$

The mean of Y is a linear function of X - this mean function represents the true relationship between X and Y. The parameters β_0, β_1, and σ are called the parameters of the regression model. They can be estimated from the data.

Example: Check out the R simulation for simple linear regression posted on Canvas. In a regression application, there is an assumed true relationship between the predictor X and the response Y. It is represented by the model parameters β_0, β_1, and σ. The data are viewed as a sample taken independently and at random from the set of all possible measures of the relationship. For a specific data set, we can estimate the regression parameters. For different data (still representing the same relationship), we would get different estimates. Regression modeling is about using the estimates obtained from the data to make statistical statements about the true model parameters.



Least-Squares Estimation of the Parameters

There are several plausible ways in which a "best-fit" regression line could be fit to a scatterplot of data that exhibits a linear relationship. By far the most common way is the "least-squares" method. The line is fit to minimize the sum of squared residuals.

Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Then the sum of squared residuals may be written as

SSRes

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Example: Minimize the quantity $S(\beta_0, \beta_1)$ with respect to β_0 and β_1.

For computations by hand, it is convenient to first compute the intermediate quantities

S_xx is constant

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

sample variance of x is: $\frac{\sum (x_i - \bar{x})^2}{n-1}$

and

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

With this, the regression parameter estimates become

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$ those 2 parameter are random variable

After the estimates of the regression slope and intercept have been obtained, the residuals can be computed as

$$\epsilon_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), \quad i = 1, \ldots, n$$

Recall, that the regression model places assumptions on the distribution of the residuals. Once the residuals have been computed (after a regression model is fit), these assumptions can be checked. We will discuss later how that is done in detail.

Example: The Rocket Propellant Data

A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together in a metal housing. The shear strength of the bond between the two types of propellant is an important characteristic. It is suspected that the shear strength (in psi) is related to the age (in weeks) of the batch of sustainer propellant. Twenty observations on shear strength (the response) and age of propellant (the predictor) are available in the file "RocketPropellant.txt" which you can download from the class website.

To read data into R, change your working directory to the directory where you saved the data (or specify a path) and type

```
rocket <- read.table("RocketPropellant.txt", header = TRUE)
```

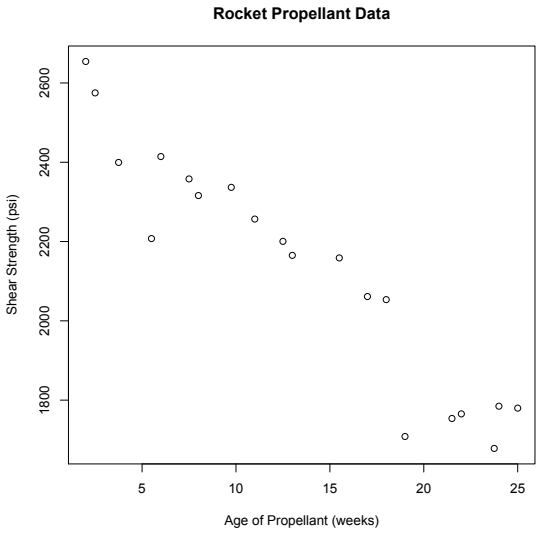
Now, `rocket` is the name of the data object in R. Type `rocket` to look at the object. Let's clean up a little and remove the observation numbers, which we don't really need and rename the variables to something more convenient:

```
rocket <- rocket[,-1]
names(rocket) <- c("strength", "age")
```

Now your data object should look like the table shown below. A good first step in every regression analysis is to draw a scatterplot of the data to convince yourself that there is really a (somewhat) linear relationship between the two variables.

```
plot(rocket$age, rocket$strength)
```

```
> rocket
   strength   age
1   2158.70 15.50
2   1678.15 23.75
3   2316.00  8.00
4   2061.30 17.00
5   2207.50  5.50
6   1708.30 19.00
7   1784.70 24.00
8   2575.00  2.50
9   2357.90  7.50
10  2256.70 11.00
11  2165.20 13.00
12  2399.55  3.75
13  1779.80 25.00
14  2336.75  9.75
15  1765.30 22.00
16  2053.50 18.00
17  2414.40  6.00
18  2200.50 12.50
19  2654.20  2.00
20  1753.70 21.50
```



Example: (cont.)

Compute the estimates of the regression intercept and slope "by hand" using the method outlined above. Of course it's ok to use software (R, Excel etc.) to compute the sums and sums of squares you'll need for this computation.

$\sum x_i = 267.25$ $\sum y_i = 42627.15$

$\sum x_i \cdot y_i = 528{,}492.64$ $\sum x_i^2 = 4677.688$

$\sum y_i^2 = 92547433.46$

$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 1{,}106.35$

$S_{xy} = -41{,}112.65$

$\beta_0 = 2627.83$

$\beta_1 = -37.15$

The same computations can also be very efficiently carried out with R:

```
> fit <- lm(strength ~ age, data = rocket)
> fit

Call:
lm(formula = strength ~ age, data = rocket)

Coefficients:
(Intercept)          age
    2627.82       -37.15
```

Properties of Least-Squares Estimators

The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have several important properties that we will make use of later when testing hypotheses. Note, that both parameter estimates can be written as linear combinations of the observations y_i.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i, \quad \text{where} \quad c_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

Since we are assuming the predictors x to be fixed and the observations y to be normally distributed, this means that the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are also normally distributed. Further, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

Example: Show that $\hat{\beta}_1$ is unbiased. $E(\hat{\beta}_1) = \beta_1$

In a similar manner, we can also compute the variance of the estimated regression parameters.

Theorem: Gauss-Markov Theorem

In a simple linear regression model in which the residuals are uncorrelated, with mean zero and equal variances the best linear unbiased estimate is given by the least squares estimates. Best here means that the least squares parameters have the smallest variance when compared with all all other unbiased estimators that are linear combinations of the y_i.

Other facts about least-squares fit regression models:

- The residuals always sum to zero.
- The sum of the observed values is the same as the sum of the fitted values.

- The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.
- The sum of the residuals weighted by the regressor variable also equals zero.

$$\sum_{i=1}^{n} x_i \epsilon_i = 0$$

- The sum of the residuals weighted by the corresponding fitted response values equals zero.

$$\sum_{i=1}^{n} \hat{y}_i \epsilon_i = 0$$

The slope and the intercept are only two of the three parameters of a simple linear regression model. We need an estimate of the residual variance σ^2 to conduct hypothesis tests and to provide confidence intervals for predictions. The estimate is usually obtained through the RESIDUAL or ERROR SUM OF SQUARES.

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

The sum of squares has $n-2$ degrees of freedom. A degree of freedom of a parameter estimate is the number of observations that go into an estimate that could be varied (without changing the estimate). It is obtained as the number of observations (here n) minus the number of parameters that have to be computed as intermediate steps. In order to compute our residuals, we need the $\hat{y}_i$. For those we first need to compute the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.

An unbiased estimate for the residual variance is

SS = sum of square
MS=mean of square

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

degree of freedom

MS here stands for "mean-square". It is a general fact in simple linear regression that

total variation of the data

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

How one partition

where SS_T is the total sum of squares which can be computed as

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

Example: Find the residual variance estimate for the Rocket Propellant data example from the R output.

Hypothesis Testing on the Slope and Intercept

Since the estimates of the regression parameters (slope, intercept and residual variance) are computed from the (random) observations (x_i, y_i), they are random variables themselves. Different data would likely have lead to slightly different regression parameter estimates. We are interested in testing hypotheses or formulating confidence intervals for the true (but unknown) regression parameters β_1, β_0, and σ^2.

Example: For instance, one might want to test, whether the regression slope is equal to some fixed value β^* against the two-sided alternative

$$H_0 : \beta_1 = \beta^* \quad \text{vs.} \quad H_a : \beta_1 \neq \beta^*$$

We have seen previously that $\hat{\beta}_1$ is normally distributed with mean β_1 and variance σ^2/S_{xx}. Thus

$$Z = \frac{\hat{\beta}_1 - \beta^*}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

by process of z-test

could be used as a test statistic. In practice, this test statistic is not very useful. Why?

Because σ is usually unknown, which is needed to calculate
The test stastic value is unknown

Which test statistic function would you suggest instead?

t-test
Replace the σ with its estimate S^2(σ ^2)=MSres, the best estimate for σ ^2

Note: The estimate of the intercept $\hat{\beta}_0$ has mean β_0 and standard error

$$\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

These two quantities can be used to construct hypothesis tests and confidence intervals for the intercept.

The most common hypothesis test in the simple linear regression scenario tests, whether the regression slope is equal to zero. What can you say about the relationship between the X and Y variables if that is the case?

If the regression slope is zero, then X cannot be used to linearly predict Y, but that doesnt mean that these could not be non-linear relationship between X and Y

Example: Construct a 95% confidence interval for the regression slope in the Rocket Propellant example. Also, test (at significance level $\alpha = 0.05$) whether the slope is equal to zero.

ANOVA Approach to Regression

ANOVA stands for analysis of variance. The idea behind ANOVA is to partition the variation in the observed response variable Y into several sources: variation that can be explained through the changing levels of the predictor X (the regression component) and variation that cannot be explained through the regression (residual variance). We have already encountered some of these components.

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

is the total sum of squares. This quantity describes the total variation (sum of squared deviations from the mean) of the response. SS_{Res} is the residual sum of squares. Recall that

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

SS_R is called the REGRESSION SUM OF SQUARES, where

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$$

It is always true for regression models that

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

or, for short

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

Each sum of squares has a corresponding degree of freedom. The degree of freedom of SS_T is $n-1$, because of the constraint $\sum(y_i - \bar{y}) = 0$. The degree of freedom of SS_R is 1, because it depends on only the regression parameter $\hat{\beta}_1$. We have seen previously, that the residual degree of freedom is $n-2$.

These ANOVA parameters are usually reported in form of a table for a regression analysis. Regardless of which software is used for the analysis, the table almost always has a format very similar to the one below. The order of columns may be slightly different in different software applications (R, for instance, lists first df, then SS, then MS)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_{Res}
Residual	$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Example: For the Rocket Propellant Data, R produces the following ANOVA table

```
> fit <- lm(strength~age, data = rocket)
> anova(fit)
Analysis of Variance Table

Response: strength
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 1527483 1527483  165.38 1.643e-10 ***
Residuals 18  166255    9236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-value column of the ANOVA table is the test statistic (sometimes with p-value) of the F-test that tests the hypothesis $H_0 : \beta_1 = 0$. In multiple regression, this null hypothesis is extended to more than one slope. In simple linear regression, this is the exact same hypothesis that we have previously tested with a t-test. In simple linear regression problems, the outcomes of these two tests (and their p-values) will always be exactly the same.

Interval Estimation in Simple Linear Regression

We have seen previously how confidence intervals for the slope and intercept parameters can be computed in simple linear regression problems. We can also compute confidence intervals for the error variance and for predicted response values at a certain predictor level x_0.

If the regression model assumptions are satisfied, it can be shown that

$$(n-2)\frac{MS_{Res}}{\sigma^2} \sim \chi^2(df = n-2)$$

Thus, to construct a confidence interval for σ^2 at level $1-\alpha$, we use the $\alpha/2$ and $1-\alpha/2$ quantiles of the χ^2-distribution with $n-2$ degrees of freedom.

$$\text{CI}_{\sigma^2}^{(1-\alpha)} = \left[\frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}, \frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}}\right]$$

Example: Find a 95% confidence interval for the residual variance σ^2 in the Rocket Propellant Example.

One important purpose of a regression model is to estimate the mean response $E(y|x_0)$ for a given predictor level x_0. Our "best guess" is of course the value from the regression line. But since the line itself is subject to uncertainty, so is this mean estimate. Note that,

$$E(y|x_0) = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

To construct a confidence interval for the true $\mu_{y|x_0}$, we need to know the distribution of the sample statistic $\hat{\mu}$. Note, that $\hat{\mu}$ is a linear combination of the $\hat{\beta}$'s, and as such normally distributed. The variance of $\hat{\mu}$ is

$$Var(\hat{\mu}_{y|x_0}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\bar{y} + \hat{\beta}_1(x_0 - \bar{x}))$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

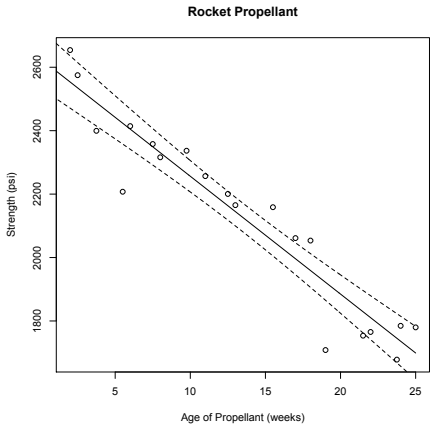
Thus, the sampling distribution of the standardized quantity is

$$\frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t(df = n - 2)$$

From there, it is easy to construct confidence intervals. Note, that the width of the confidence interval depends on the quantity $(x_0 - \bar{x})^2$ in the denominator. Close to $\bar{x}$, the prediction of the mean is more precise than far away from the mean where we have less data to make predictions.

Example: For the Rocket Propellant data, we can create a 95% confidence interval for the mean response with the following code:

```
> model <- lm(strength~age, data = rocket)
> plot(rocket$age, rocket$strength, xlab = "Age of Propellant (weeks)", ylab = "Strength (psi)", main = "Rocket Propellant")
> xv <- seq(0,25,0.1)
> yv <- predict(model, list(age = xv), int = "c", level = 0.95)
> matlines(xv, yv, lty=c(1,2,2), col = "black")
```



Here, the dotted lines show the bounds of the 95% confidence interval. Note, that the intervals flare for age values that are further away from the mean.

Prediction of New Observations

Predicting the mean of new observations is fundamentally different than predicting a single new observation. Not only is there uncertainty about predicting the regression line based on a sample of observations, but there is additional uncertainty about predicting where (on, above, or below...) the line the new observation may fall. Thus, confidence intervals for single predictions are wider than confidence intervals for the mean (if the level α is held fixed.)

Consider a confidence interval for y_0, where

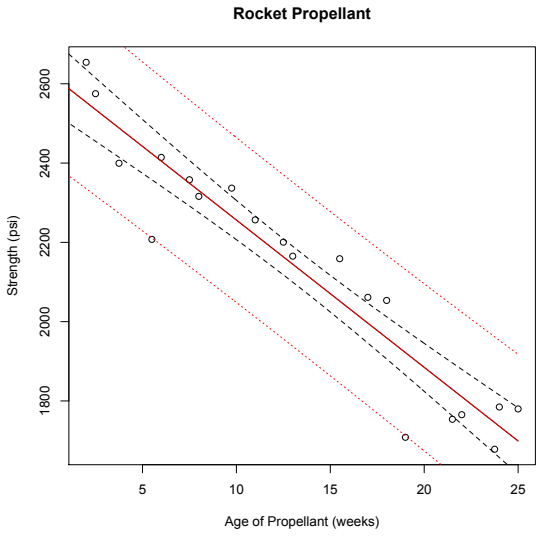
$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

The mean of y_0 is, of course, $\hat{y}_0$. Let $\psi = y_0 - \hat{y}_0$. Then ψ is normally distributed (since y_0 is normal and $\hat{y}_0$ is a linear combination of the $\hat{\beta}$'s.) And $E(\psi) = 0$ and

$$Var(\psi) = Var(y_0 - \hat{y}_0) = Var(y_0) + Var(\hat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Example: For the Rocket propellant data, we can compute a 95% prediction interval for a new observation with the following code. Note, that the `int = "p"` specifies computing the lower and upper bound for a prediction interval (as opposed to `"c"` for a confidence interval for the mean). Note further that the prediction interval for a new observation (shown in red) is wider than the confidence interval for the mean (in black). The prediction interval also has a slight bend (which is less noticable than the bend in the confidence interval for the mean).

```
> model <- lm(strength~age, data = rocket)
> plot(rocket$age, rocket$strength, xlab = "Age of Propellant (weeks)", ylab = "Strength (psi)", main = "Rocket Propellant")
> xv <- seq(0,25,0.1)
> yv <- predict(model, list(age = xv), int = "c", level = 0.95)
> zv <- predict(model, list(age = xv), int = "p", level = 0.95)
> matlines(xv, yv, lty=c(1,2,2), col = "black")
> matlines(xv, zv, lty=c(1,3,3), col = "black")
```



Coefficient of Determination

The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

is called the COEFFICIENT OF DETERMINATION. R^2 can be interpreted as the percentage of variation in the response Y that is explained through the regression on the predictor X. Since $0 \leq SS_R \leq SS_T$, the coefficient of determination is always $0 \leq R^2 \leq 1$. Values close to 1 imply that the regression can explain most of the variation in Y. Values close to 0 imply that there is no linear relationship between X and Y. In simple linear regression problems, R^2 is simply the square of the correlation coefficient between X and Y. In multiple linear regression, R^2 can still be interpreted as the percentage of variation in the response explained through the regression on now multiple predictors.

The coefficient of determination can be used as one tool to decide whether a regression model is a good fit. But it should not be used as the only tool. For instance, if there is a very good non-linear relationship between X and Y, the value of R^2 may still be very low. In general, the magnitude of R^2 depends on the number of predictor variables in the model. If predictors are added, the value of R^2 will always increase. In addition, R^2 increases if the spread of the predictor variable(s) increases.

Regression Through the Origin

Sometimes the nature of the problem suggests that the regression line should go through the origin.

Example: Income earned as a function of time worked.

The statistical model for regression through the origin is

$$y = \beta_1 x + \epsilon$$

Example: Find the least squares estimate of the slope β_1 in this model.

The estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

is unbiased. The estimate of the residual variance σ^2 has n-1 degrees of freedom

$$\hat{\sigma}^2 = \text{MS}_{Res} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} y_i x_i}{n-1}$$

Similarly to the regular simple linear regression case, confidence intervals for the slope parameter β_1 and for the mean response or a single new observation can be constructed based on the assumption that the errors are normally distributed.

To decide whether to fit a regression model with or without intercept to a specific application, a scatter plot can be helpful. If most observations have been taken far away from the origin, observing the scatter plot may not be enough. When the linear regression model with intercept is fit, we can test the hypothesis $H_0 : \beta_0 = 0$. When this hypothesis cannot be rejected, a model without intercept may be more appropriate. The residual mean square is a good way to compare models with and without intercept. The better fitting model should have smaller (squared) residuals. R^2 is not a good way to compare the fit of linear regression models with and without intercept. While R^2 measures the proportion of variation in y (around $\bar{y}$) in the regular simple linear regression model, the corresponding statistic in the linear regression model through the origin measures the proportion of variability around the origin.

To fit a linear model of y as a function of x without intercept in R use the command `y ~ 0 + x`.

Cases where the Regressor x is Random

So far we have assumed that the values of the regressor X are determined (without error) by the experimenter. There are many situations where this assumption is inappropriate, as both X and Y are related but random quantities that are being observed. As long as the conditional distribution of Y given X can be assumed to be Normal$(\beta_0 + \beta_1 x, \sigma^2)$, the formulas that we have previously derived are still valid.