Contents

1 Simple Linear Regression

The linear regression model is applied if we want to model a numeric response variable and its dependency on at least one numeric predictor variable.

The Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + e, \quad e \sim \mathcal{N}(0, \sigma)$$

where

y = dependent or response variable

x = independent or predictor variable

e = random error, assumed to be normally distributed with mean 0 and standard deviation σ

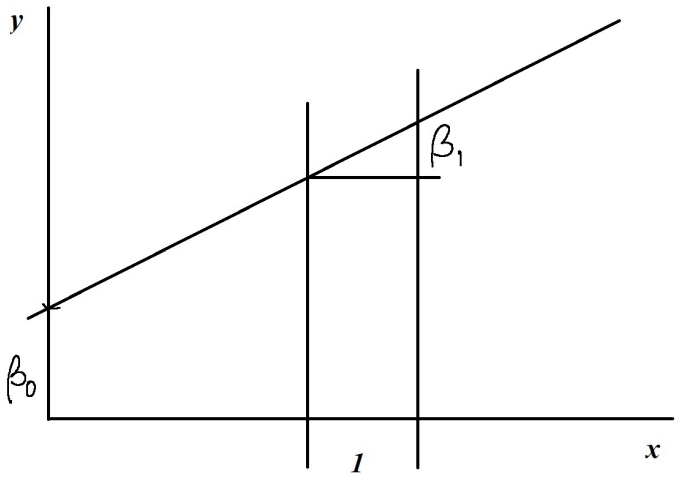
β_0 = intercept

β_1 = slope

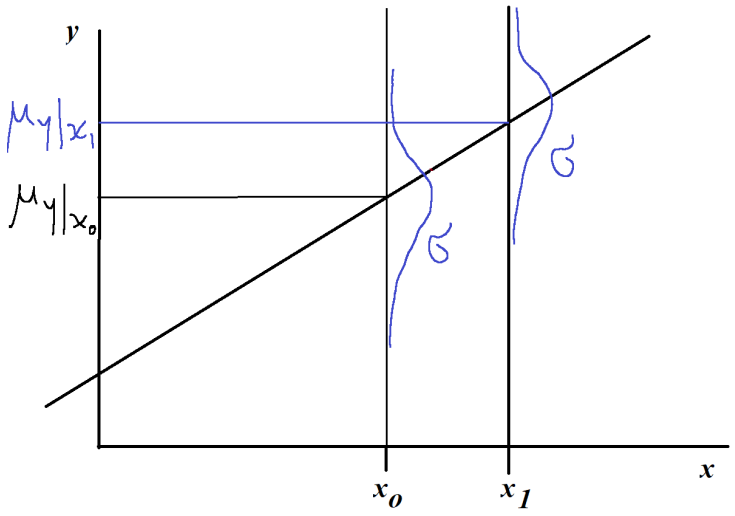
A consequence of this model is that for a given value of x the mean of the response variable y follows a line in the predictor variable, x.

$$\mu_y = E(y) = \beta_0 + \beta_1 x$$

$\beta_0 + \beta_1 x$ describes a line with y-intercept β_0 and slope β_1



In conclusion the model implies that y on average follows a line, depending on x.
Since y is assumed to also underlie a random influence, not all data is expected to fall on the line, but that the line will represent the mean of y for given values of x.
The randomness in y is captured through the error, e, which describes that measurements will be scattered around the line. To be more precise, once we fix a value of $x = x_0$, then y is on average equal to $\beta_0 + \beta_1 x_0$ and has a standard deviation of σ (the standard deviation of the error, e).



We will see how to use sample data to estimate the parameters of the model, how to check the usefulness of the model, and how to use the model for prediction, estimation and interpretation.

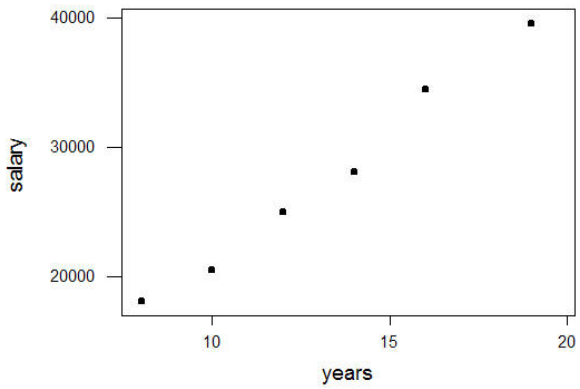
1.1 Fitting the Model

In a first step obtain a scattergram, to check visually if a linear relationship seems to be reasonable.

Example 1
Does the number of years invested in schooling pay off in the job market?
Apparently so – the better educated you are, the more money you will earn. The data in the following table give the median annual income of full-time workers, age 25 or older by the number of years of schooling completed - this is old data, newer data can be found in the example at the end of these notes.

x=Years of Schooling	y=Median Income (in 1000 dollars)
8	18
10	20.5
12	25
14	28.1
16	34.5
19	39.7

Start of with creating a scatterplot for x and y.



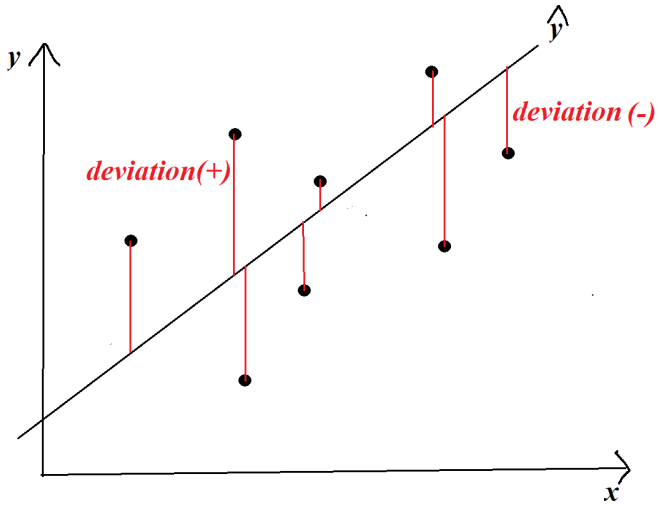
The scatterplot shows a strong, positive, linear association between years of schooling and the median salary. Indicating that a SLR model should be appropriate for describing the relationship between the two variables.

Questions to be addressed based on the scatterplot:

1. Can the relationship between x and y be described by a straight line?
   If yes, what is the nature of the relationship? (positive or negative)
2. Is there a non-linear relationship between x and y?
3. How strong is the association? (weak, moderate, strong)

If the relationship between two variables is linear and has no random component, the dots in the scatterplot would all fall precisely on a line and the deviations from the line would all be 0. But if there is a random component, the data in the scatterplot will exhibit a linear pattern that generally follows a strait line, with some of the measurements being close and others further away from the line.
When interpreting the line as the prediction of y given values of x, we define

deviations = difference between observed measurements and the line = error of prediction



The line that results in the least squared deviations (for the sample data) will be used as estimation for the line $\beta_0 + \beta_1 x$ in the model (population).

Continue example:
We propose a SLR model for the association between income (y) and years of schooling (x):

$$y = \beta_0 + \beta_1 x + e, \quad e \sim \mathcal{N}(0, \sigma)$$

After we established through the scatterplot that a linear model seems to be appropriate for describing the association between x and y, in a next step the estimates for the model parameters are calculated using the given sample data, by making the total squared deviations between line and measurements as small as possible. The result will be the estimated regression line

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

The hats indicate that these are estimators for the true population parameters, based on sample data.

For given $\hat{\beta}_0$ and $\hat{\beta}_1$, the deviation of the ith value from its predicted value is $y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.
Then the sum of squared deviations for all n data points is

$$SSE = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$\hat{\beta}_0$ and $\hat{\beta}_1$, so that the SSE is as small as possible, and call the result the "least squares estimators" of the population parameters β_0 and β_1.

Solving this problem mathematically provides:
With

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \quad \text{and} \quad SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

- the estimator for the slope is $\hat{\beta}_1 = \dfrac{SS_{xy}}{SS_{xx}}$ and
- the estimator for the intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Continue example:
To apply the formulas first get the required totals from the data:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	8	18	64	324	144
	10	20.5	100	420.25	205
	12	25	144	625	300
	14	28.1	196	789.61	393.4
	16	34.5	256	1190.25	552
	19	39.7	361	1576.09	754.3
totals	79	165.8	1121	4925.2	2348.7
	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$

Using these totals calculate:

$$\begin{aligned} SS_{xy} &= 2348.7 - \frac{(79)(165.8)}{6} &&= 165.667 \\ SS_{xx} &= 1121 - \frac{(79)^2}{6} &&= 80.833 \\ \hat{\beta}_1 &= \frac{165.667}{80.833} &&= 2.049 \\ \hat{\beta}_0 &= \frac{165.8}{6} - 2.049\frac{79}{6} &&= 0.648 \end{aligned}$$

So,

$$\hat{y} = 0.648 + 2.049x$$

describes the line that results in the least total squared deviations (SSE) possible, and is called the *Least Square Regression Line/Equation*.

The intercept ($\hat{\beta}_0$) estimates that the mean income is \$648 for a person with zero years of education.
Using this interpretation would mean that we apply the result outside the range of observed values of x. This is called extrapolation, which is highly problematic and usually not admissible.

The slope ($\hat{\beta}_1$) estimates that in average the income increases by \$2049 for every extra year of schooling.

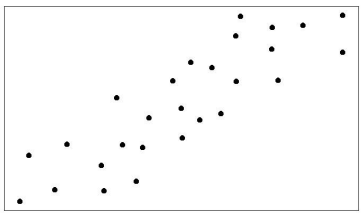
1.2 Understanding the Model

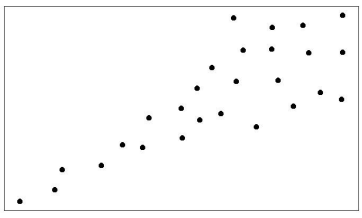
- The first implication of the model is that the mean of y follows a straight line in x, not a curve

  Write $\mu_y = E(y) = \beta_0 + \beta_1 x$

- The model implies that the deviations from the line are normally distributed (the error is normally distributed)

- The variation from the line is assumed to be the same across the range of x (σ, the standard deviation of the error is the same for all values of x). This is called the homoscedasticity assumption





The data in the figure on the left do not indicate a violation of the homoscedasticity assumption, the scatter is the about the same across the values of x, but in the figure on the right the scatter increases as x increases indicating a violation of the homoscedasticity assumption.

- The measurements in the sample are independent

This discussion shows that another important parameter in the model (beside β_0 and β_1) is the standard deviation of the error, σ.

Estimating σ
A large standard deviation indicates that the deviations, i.e. the difference between the line and the measured values, are large on average. Therefore it makes sense that $SSE = \sum(\hat{y}_i - y_i)^2$ is used in the estimator for σ.
In fact the best estimator for σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2}$$

where

$$SSE = \sum(\hat{y}_i - y_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

and

$$SS_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

and the best estimator for σ is

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

It is called the estimated standard error of the regression model.

Continue example:
Find the estimate for the standard deviation σ from the model, which explains how far the data points fall from the line.

$$\begin{aligned} SS_{yy} &= 4925.2 - \frac{(165.8)^2}{6} = 343.593 \\ SSE &= 343.593 - 2.049(165.667) = 4.142 \end{aligned}$$

then

$$s = \sqrt{\frac{4.142}{6-2}} = 1.02$$

is the estimated standard deviation in the regression model.
Interpretation of s:
One should expect about 95% of the observed y values to fall within $2s$ of the estimated line.
We expect that 95% of the median incomes fall within $2s = \$2.04$ from the true population regression line.

1.3 Measuring Model Fit

Correlation Coefficient

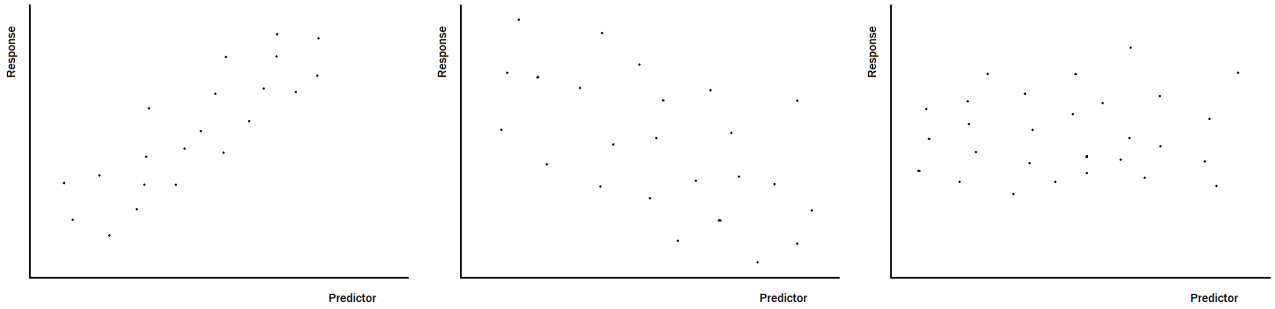
The correlation coefficient

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

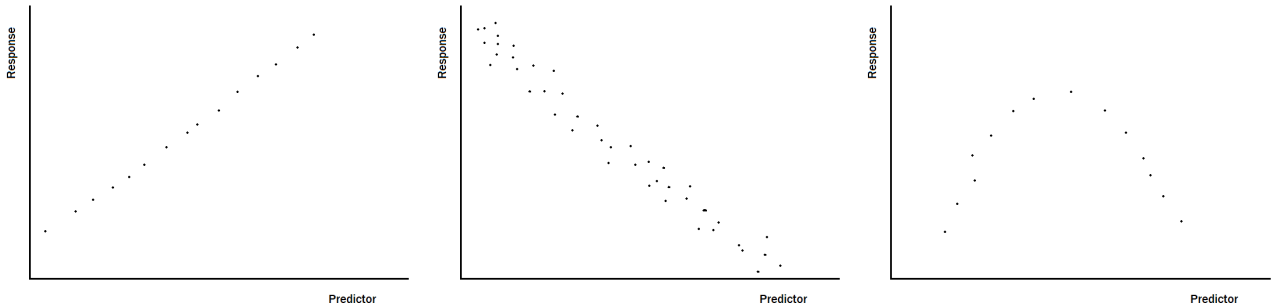
measures the strength of the linear relationship between two variables x and y.

Properties

1. $-1 \leq r \leq 1$
2. independent from units used
3. $abs(r) = 1$ indicates that all values fall precisely on a line
4. $r > (<)0$ indicates a positive (negative) relationship.
5. $r \approx 0$ indicates no linear relationship



$r \approx 0.8$ $r \approx -0.5$ $r \approx 0$



$r \approx 1$ $r \approx -0.9$ $r \approx 0$(nonlinear)

Continue example:

$$r = \frac{165.667}{\sqrt{80.833(343.593)}} = 0.994$$

indicates a very strong positive linear relationship between the years of schooling and the mean income.

Coefficient of Determination

The coefficient of determination measures the usefulness of the model, it is the correlation coefficient squared, r^2, and a number between 0 and 1. It's wide use can be explained through it's interpretation.

An alternative formula for r^2 is

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

This formula can help to understand the interpretation of r^2.

SS_{yy} measures the spread in the y values (in the sample), and SSE measures the spread in y we can not explain through the linear relationship with x. So that $SS_{yy} - SSE$ is the spread in y explained through the linear relationship with x. By dividing this by SS_{yy}, we find that r^2 gives the proportion in the spread of y, that is explained through the linear relationship with x.

Continue Example:
The coefficient of determination in our example is

$$r^2 = 0.994^2 = 0.988$$

98.8% in the variation of the mean income is explained through years of schooling.
A high coefficient of determination indicates that the model is suitable for estimation and prediction.

1.4 Checking Model Utility – Making Inference about the slope

As the sample mean $\bar{x}$ is an estimate for the population mean μ, the sample slope $\hat{\beta}_1$ is an estimate for the population slope β_1.
Our next goal is to find a confidence interval for β_1, and to learn how to conduct a test concerning the true population slope, β_1.

In order to this we have to discuss **the distribution of** $\hat{\beta}_1$.
If the assumptions of the regression model hold, then $\hat{\beta}_1$ is normally distributed with mean $\mu_{\hat{\beta}_1} = \beta_1$ and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

$\sigma_{\hat{\beta}_1}$ is estimated by

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

which is the estimated standard error of the least squares slope $\hat{\beta}_1$.

Combining all this information we find that under the assumption of the regression model

$$t = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SS_{xx}}}$$

is t-distributed with $df = n - 2$.

$(1 - \alpha) \times 100\%$ **CI for** β_1

$$\hat{\beta}_1 \pm t^* \frac{s}{\sqrt{SS_{xx}}}$$

where $t^* = (1 - \alpha/2)$ percentile of the t-distribution with $df = n - 2$.

Continue example:
A 95% CI for β_1 (df=4)

$$2.049 \pm (2.776)\frac{1.02}{\sqrt{80.833}} \rightarrow 2.049 \pm 0.312$$

A 95% CI for β_1 is [1.737 ; 2.361]. We are 95% confident that the mean income increases between \$1738 and \$2360 for every extra year of schooling.

Does the CI provide sufficient evidence that $\beta_1 \neq 0$?
Since zero is not included with the confidence interval, we are 95% confident that the slope is not zero,i.e. $\beta_1 \neq 0$.
This is an important question, because, if β_1 would be 0, then the variable x would contribute no information on y using a linear model.

Another way of answering this question is the **Model Utility Test about** β_1

1. Hypotheses: Parameter of interest β_1.

test type	hypotheses
upper tail	$H_0 : \beta_1 \leq 0$ vs. $H_a : \beta_1 > 0$
lower tail	$H_0 : \beta_1 \geq 0$ vs. $H_a : \beta_1 < 0$
two tail	$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

   Choose α.

2. Assumptions: Regression model is a proper description of the population.

3. Test statistic:
$$t_0 = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}, \quad df = n - 2$$

4. P-value/Rejection Region:

test type	P-value	Rejection Region
upper tail	$P(t > t_0)$	$t_0 > t_\alpha$
lower tail	$P(t < t_0)$	$t_0 < t_\alpha$
two tail	$2P(t > abs(t_0))$	$abs(t_0) > t_{\alpha/2}$

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Put into context

Continue example:
Test if the data provide sufficient evidence that the mean income increases with increasing years of schooling.
Test if $\beta_1 > 0$

1. Hypotheses: Parameter of interest β_1.

   $H_0 : \beta_1 \leq 0$ vs. $H_a : \beta_1 > 0$ Choose $\alpha = 0.05$.

2. Assumptions: Regression model is a proper description of the population.

3. Test statistic:
$$t_0 = \frac{2.049}{1.02/\sqrt{80.833}} = 18.240, \quad df = 4$$

4. P-value: upper tail $P(t > t_0) < 0.0001$

5. Decision: $P - value < 0.0001 < 0.05 = \alpha$ reject H_0

6. At significance level of 5% we conclude that the mean income increases if the number of years of schooling increases. The linear regression model helps explaining the mean income.

1.5 Applying the Model – Estimation and Prediction

Once we established that the model we found describes the relationship properly, we can use the model for

- estimation: estimate mean values of y, $E(y)$ for given values of x. For example we might be interested in the mean income of people who have 15 years of schooling.
- prediction: new individual value based on the knowledge of x, For example, we want to give the range of income most people with 15 years of schooling fall within.

Which application will we be able to answer with higher accuracy?
The key to both applications of the model is the least squares line, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, we found before, which gives us point estimators for both situations..
The next step is to find confidence intervals for $E(y)$ for a given x, and for an individual value y, for a given x.
In order to do this we will have to discuss the distribution of $\hat{y}$ for the two applications.

Sampling distribution of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

1. The standard deviation of the distribution of the estimator $\hat{y}$ of the mean value $E(y)$ at a specific value of x, say x_p

$$\sigma_{\hat{y}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{Sxx}}$$

   This is the standard error of $\hat{y}$.

2. The standard deviation of the distribution of the prediction error, $y - \hat{y}$, for the predictor $\hat{y}$ of an individual new y at a specific value of x, say x_p

$$\sigma_{(y-\hat{y})} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   This is the standard error of prediction.

A $(1-\alpha)100\%$ CI for $E(y)$ at $x = x_p$

$$\hat{y} \pm t^* \; s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

A $(1-\alpha)100\%$ Prediction Interval for y at $x = x_p$

$$\hat{y} \pm t^* \; s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

Comment: Do a confidence interval, if interested in the mean of y for a certain value of x, but if you want to get a range where a future value of y might fall for a certain value of x, then do a prediction interval.

Continue example:
According to the linear regression model predict the income after $x = 15$ years of schooling? Find a prediction interval.

For $x_p = 15$ find $\hat{y} = 0.648 + 2.049(15) = 31.383$. With $df = 4$ $t^* = 2.776$

$$31.383 \pm 2.776(1.01)\sqrt{1 + \frac{1}{6} + \frac{(15 - 79/6)^2}{80.833}}, \text{gives} \quad 31.383 \pm 3.082$$

which is $[28.301, 34.465]$. For a person with with 15 years of schooling with 95% confidence the income falls between \$28301 and \$34465.
Estimate the mean income for people with 15 years of schooling at confidence level of 95%.

$$31.183 \pm 2.776(1.01)\sqrt{\frac{1}{6} + \frac{(15 - 79/6)^2}{80.833}}, \text{gives} \quad 31.183 \pm 1.279$$

which is $[29.858, 32.417]$. With 95% confidence the mean income for people with 15 years of schooling falls between \$29858 and \$32417.