LINEAR REGRESSION

Linear Regression Analysis

- Regression analysis is used to predict the value of one variable (the dependent variable) on the basis of other variables (the independent variables).
  - Dependent variable: denoted Y
  - Independent variables: denoted X_1, X_2, …, X_k
  - If we only have ONE independent variable, the model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

which is referred to as simple linear regression. We would be interested in estimating β_0 and β_1 from the data we collect.

Linear Regression Analysis

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Variables:
  - X = Independent Variable (we provide this)
  - Y = Dependent Variable (we observe this)
- Parameters:
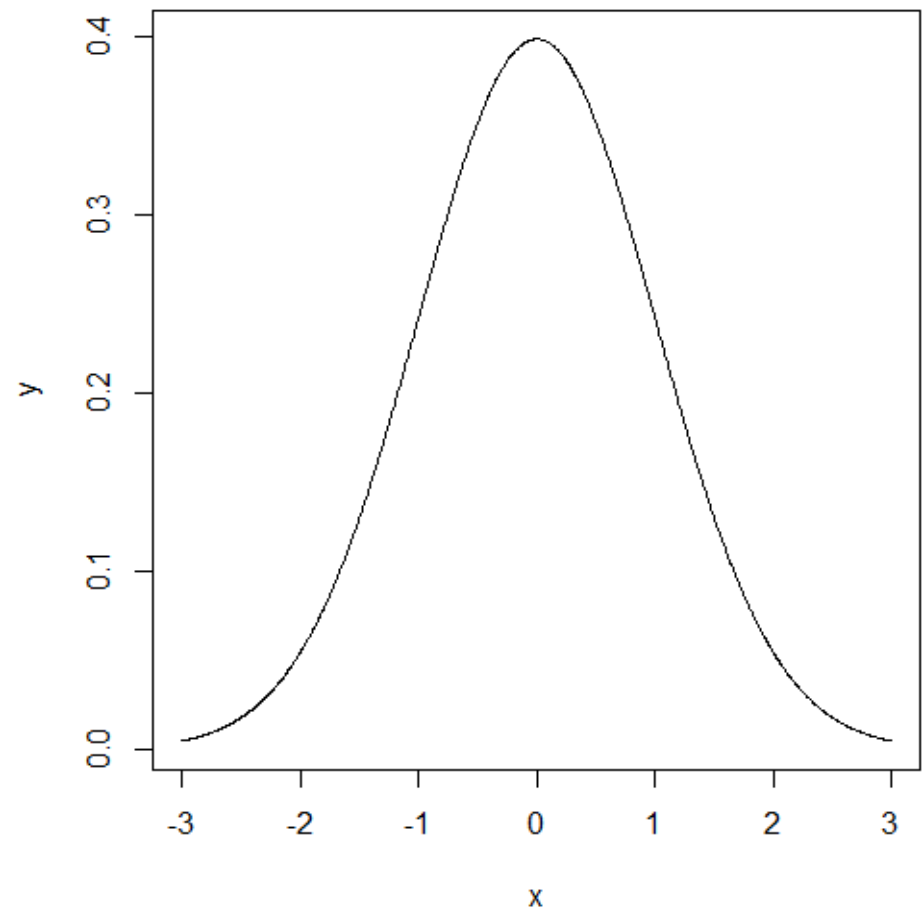 - β_0 = Y-Intercept
  - β_1 = Slope
  - $\varepsilon \sim$ Normal Random Variable ($\mu_\varepsilon = 0$, $\sigma_\varepsilon = ???$) [Noise]
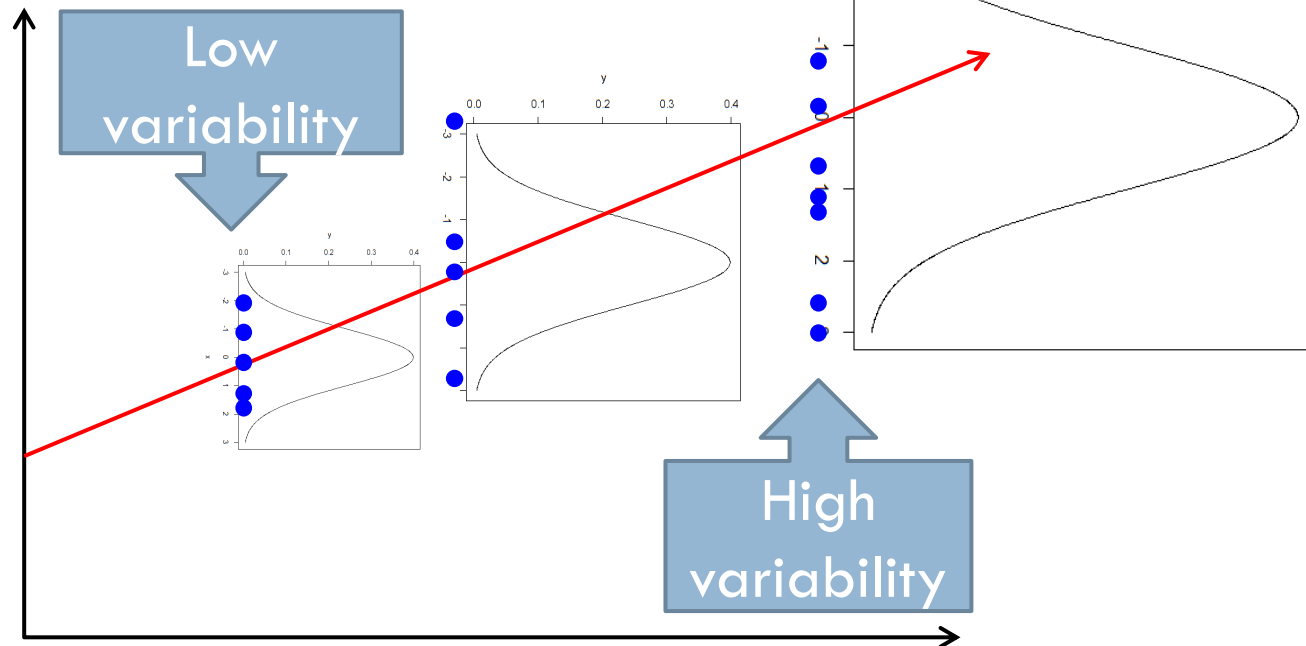
Effect of Larger Values of σ_ε

- x <- seq(-3,3,length=1000)
- y <- dnorm(x, mean=0, sd=1)
- plot(x,y, type="l", lwd=1)

Theoretical Linear Model

This model have a problem with homoscedasticity, as we review later

Low variability

High variability

Building the model

Collect Data
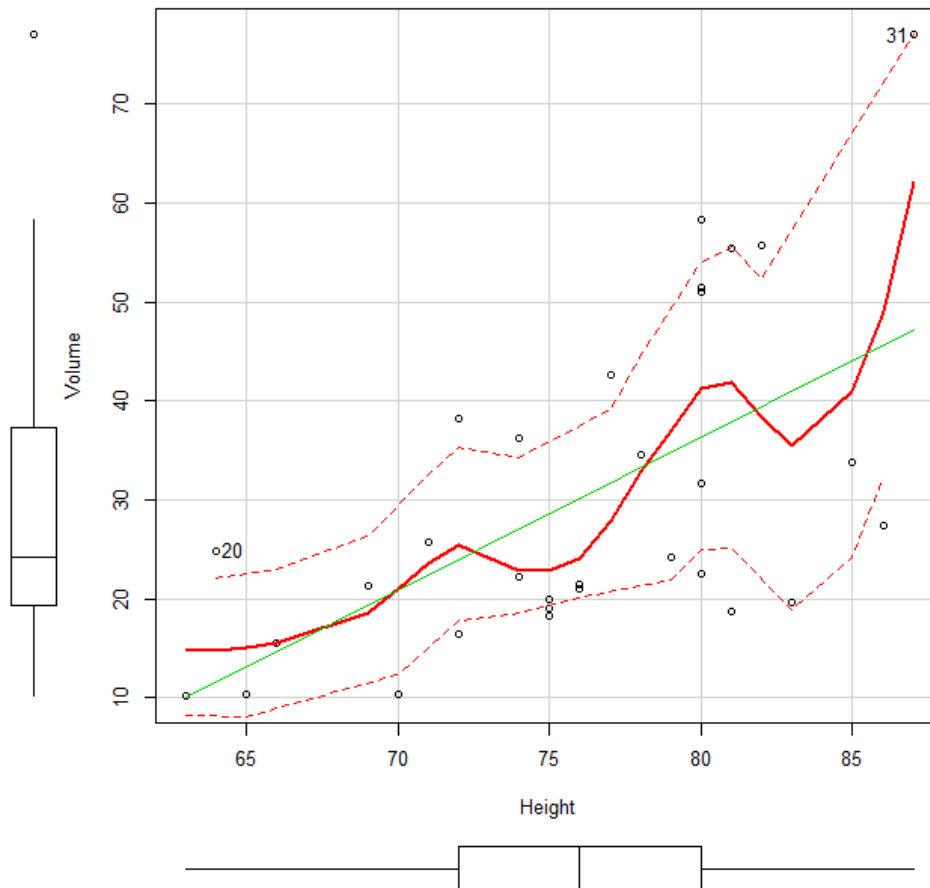
- Can we predict the result of the second test depending on the first test?

  - Test 2 Grade = β_0 + β_1*(Test 1 Grade)

- From the data:

  - Estimate β_0
  - Estimate β_1
  - Estimate σ_ε

Student	1st mark	2nd mark
1	6	7
2	7	5
..		
n	9	10

Correlation Analysis… "-1 < ρ < 1"

☐ If we are interested only in determining whether a relationship exists, we employ **correlation analysis.** (Example: trees height and volume.)

Rcmdr> cor(trees[,c("Height","Volume")], use="complete")
     Height   Volume
Height 1.0000000 0.5982497
Volume 0.5982497 1.0000000

Rcmdr> scatterplot(Volume~Height, reg.line=lm, smooth=TRUE, spread=TRUE,
Rcmdr+   id.method='mahal', id.n = 2, boxplots='xy', span=0.5, data=trees)

Correlation Analysis… "-1 < ρ < 1"

- If the correlation coefficient is **close to +1** that means you have a **strong positive relationship.**

- If the correlation coefficient is **close to -1** that means you have a **strong negative relationship.**

- If the correlation coefficient is **close to 0** that means you have **no correlation.**

- We have the ability to test the hypothesis
  - H0: ρ = 0

$$\cos(\alpha) = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$
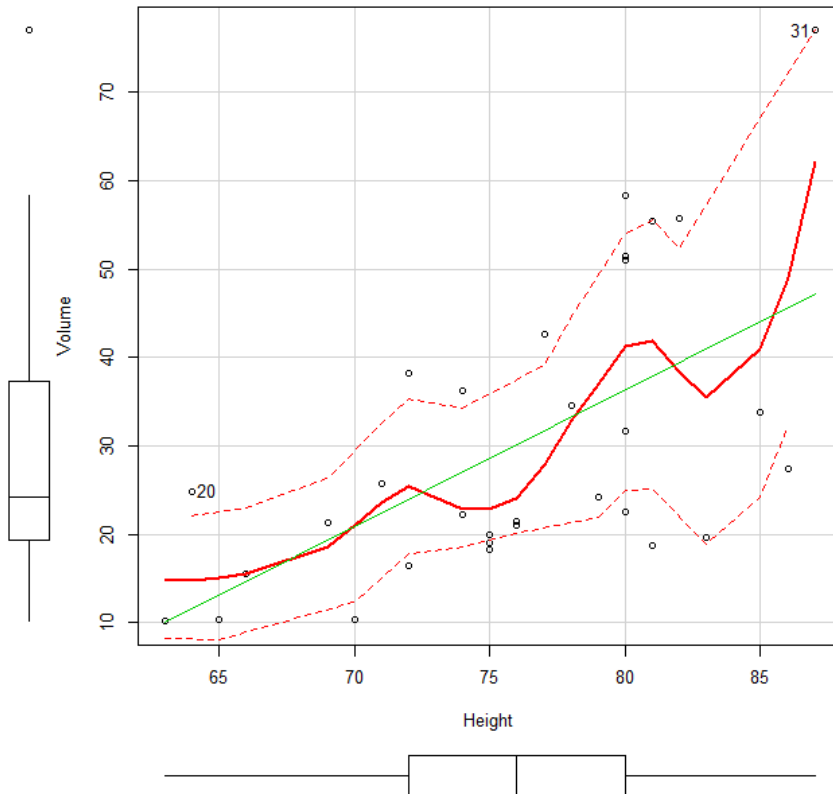
Regression

- Model Types… X=size of house, Y=cost of house
- **Deterministic Model**: an equation or set of equations that allow us to fully determine the value of the dependent variable from the values of the independent variables.
  - $y = 69 + 0.23x$
  - Area of a circle: $A = \pi \ast r2$
- **Probabilistic Model**: a method used to capture the randomness that is part of a real-life process.
  - $y = 69 + 0.23x + \varepsilon$
- Ex. do all trees of the same volume (measured in square feet) have the same height?
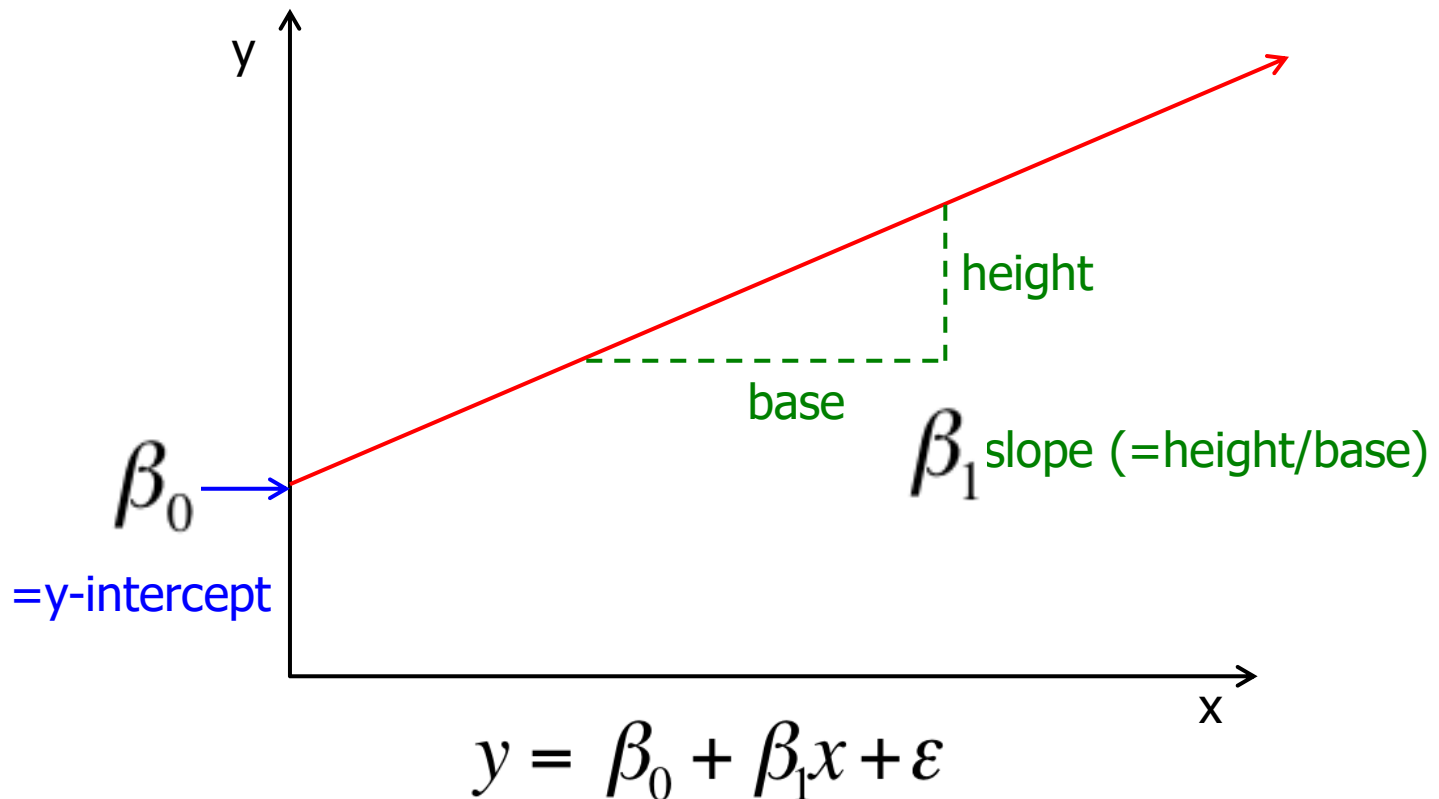
Linear Regression Analysis

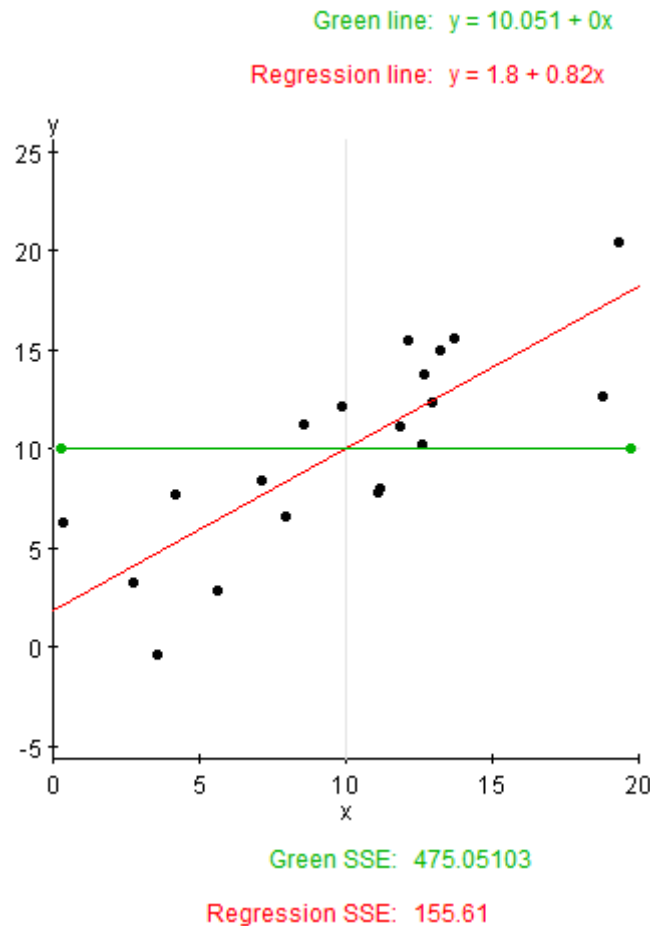- Always we can find a line that "covers" all the points.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Simple Linear Regression Model…

□ Meaning of β_0 and β_1

□ $\beta_1 > 0$ [positive slope] $\beta_1 < 0$ [negative slope]

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Plotted Regression Model

Green line: $y = 10.051 + 0x$

Regression line: $y = 1.8 + 0.82x$

Green SSE: 475.05103

Regression SSE: 155.61

☐ Play with the lines to get a "good" plot.

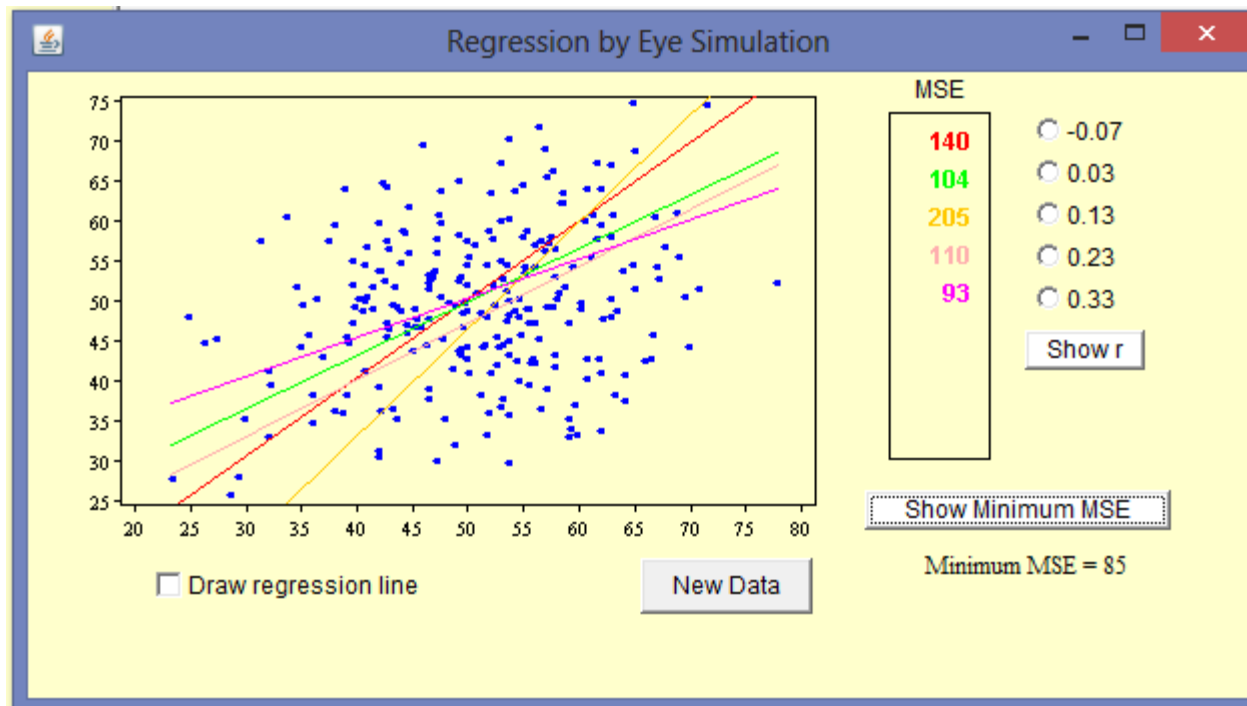☐ See: http://www.stat.tamu.edu/~west/ph/regeye.html

Which line has the best "fit" to the data?

□ **Regression applet:**

■ http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

Estimating the Coefficients

☐ In much the same way we base estimates of μ on $\bar{X}$, we estimate β_0 with b_0 and β_1 with b_1, the y-intercept and slope (respectively) of the *least squares* or *regression line* given by:

$$y = \beta_0 + \beta_1 x \qquad\qquad \hat{y} = b_0 + b_1 x$$

☐ This is an application of the least squares method and it produces a straight line that *minimizes* the sum of the squared differences between the points and the line.
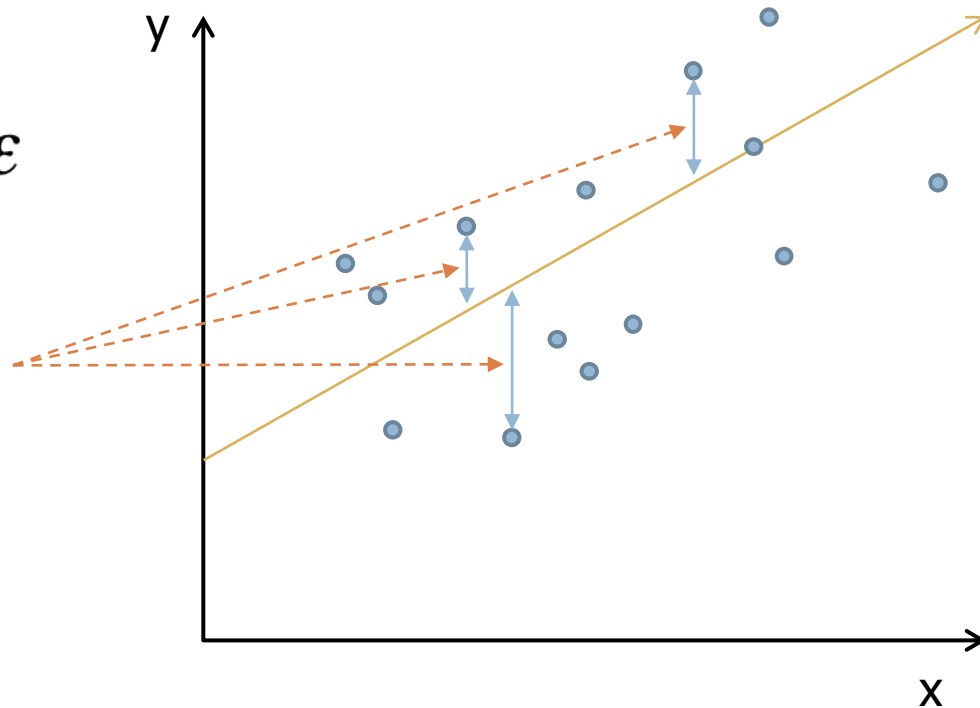
Least Squares Line

- This line minimizes the differences of the squares between the points and the line.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Residual errors

Least Squares Line

□ We start from the data

Derivation of Regression Parameters

☐ The error in the ith observation is:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

☐ For a sample of n observations, the mean error is:

$$\begin{aligned}
\bar{e} &= \frac{1}{n} \sum_i e_i = \frac{1}{n} \sum_i \{y_i - (b_0 + b_1 x_i)\} \\
&= \bar{y} - b_0 - b_1 \bar{x}
\end{aligned}$$

☐ Setting mean error to zero, we obtain:

$$b_0 = \bar{y} - b_1 \bar{x}$$

☐ Substituting b0 in the error expression, we get:

$$e_i = y_i - \bar{y} + b_1 \bar{x} - b_1 x_i = (y_i - \bar{y}) - b_1(x_i - \bar{x})$$

Derivation of Regression Parameters (Cont)

- The sum of squared errors SSE is:

$$
\begin{aligned}
\mathrm{SSE} &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} \left\{ (y_i - \bar{y})^2 + 2b_1 (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 (x_i - \bar{x})^2 \right\} \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2b_1 \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \\
&\quad + b_1^2 \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= s_y^2 - 2b_1 s_{xy}^2 + b_1^2 s_x^2
\end{aligned}
$$

Derivation (Cont)

- Differentiating this equation with respect to b_1 and equating the result to zero:

$$\frac{d(\mathrm{SSE})}{db_1} = -2s_{xy}^2 + 2b_1 s_x^2 = 0$$

- That is,

$$b_1 = \frac{s_{xy}^2}{s_x^2} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2}$$

Least Squares Line

□ The coefficients b1 and b0 for the least squares line $\hat{y} = b_0 + b_1 x$ are calculated as:

Covariance

Variance

$$s_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
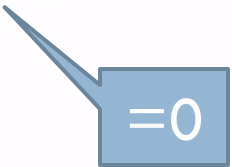
$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$\bar{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

Proof

$$\sum_{i=1}^{n}(Y_i - \mu)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y} + \bar{Y} - \mu)^2$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \bar{Y}) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)(\sum_{i=1}^{n}Y_i - n\bar{Y}) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$$

$$\geq \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

=0

Using gradient descent

- https://machinelearningmastery.com/linear-regression-tutorial-using-gradient-descent-for-machine-learning/

Least Squares Line

☐ From data, using statistics to obtain information and predict.

$$y = \beta_0 + \beta_1 x + \varepsilon$$
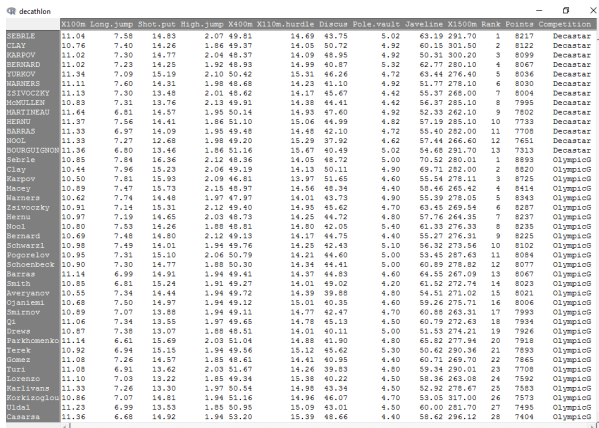
Assessing the Model

Is the model good?

Assessing the Model

☐ The least squares method will **always produce a straight line**, even if there is no relationship between the variables, or if the relationship is something other than linear.

☐ Hence, in addition to determining the coefficients of the least squares line, we need to assess it to see how well it "fits" the data. We'll see these evaluation methods now. They're based on the what is called sum of squares for errors (SSE).

Sum of Squares for Error

- The sum of squares for error is calculated as:

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right)$$

and is used in the calculation of the standard error of the estimate:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

- If S_ε is zero, all the points fall on the regression line.

Standard Error

□ If S_ε is small, the fit is excellent and the linear model should be used for forecasting. If S_ε is large, the model is poor… but what is small or large?

Rcmdr> RegModel.1 <- lm(X100m~X110m.hurdle, data=decathlon)

Rcmdr> summary(RegModel.1)

Call:

lm(formula = X100m ~ X110m.hurdle, data = decathlon)

Residuals:

   Min   1Q Median   3Q   Max

-0.44870 -0.16112 0.03165 0.11431 0.53716

Coefficients:

      Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.27614   1.06281   5.905 7.01e-07 ***

X110m.hurdle 0.32329   0.07273   4.445 7.08e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.217 on 39 degrees of freedom S_ε

Multiple R-squared: 0.3363,   Adjusted R-squared: 0.3193

F-statistic: 19.76 on 1 and 39 DF, p-value: 7.082e-05

Standard Error

- Judge the value of S_ε by **comparing** it to the **sample mean** of the dependent variable ($\bar{y}$).

- In this example,
  - $S_\varepsilon = .217$ and
  - $\bar{y} = 11.0$

- It appears to be "small", hence our linear regression model is "good".

Testing the Slope

□ If no linear relationship exists between the two variables, we would expect the regression line to be horizontal, that is, to have a slope of zero.
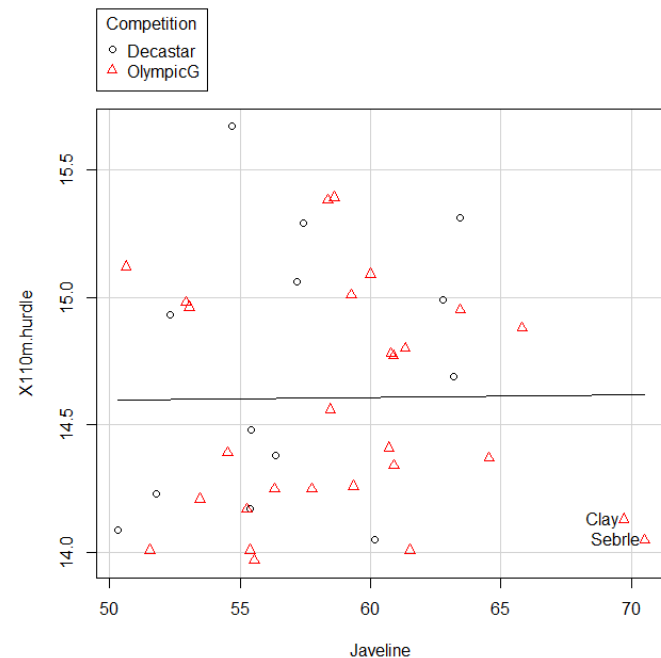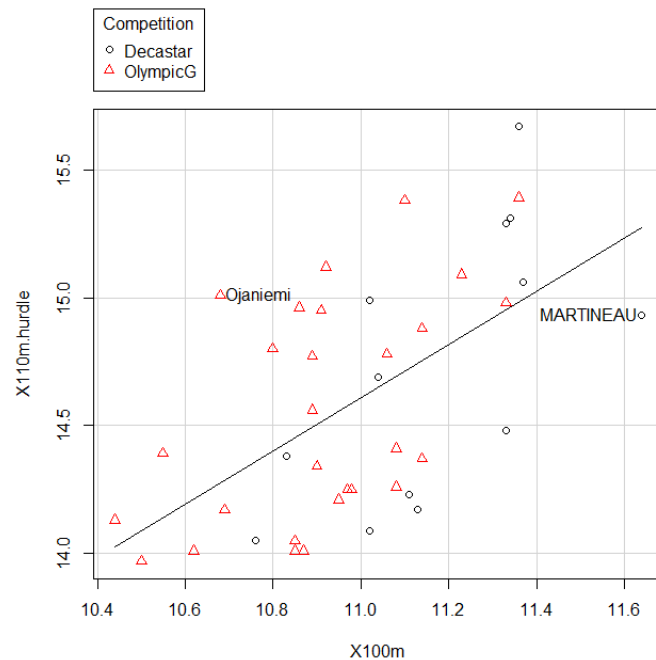
Testing the Slope

☐ We want to see if there is a linear relationship, i.e. we want to see if the slope (β_1) is something other than zero. Our research hypothesis becomes:

  ☐ H1: $\beta_1 \neq 0$

☐ Thus the null hypothesis becomes:

  ☐ H0: $\beta_1 = 0$

Testing the Slope

- We can implement this test statistic to try our hypotheses:
    - H0: $\beta_1 = 0$

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

- Where S_{b_1} is the standard deviation of b_1, defined as:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

- If the error variable (ε) is normally distributed, the test statistic has a Student **t**-distribution with n–2 degrees of freedom. The rejection region depends on whether or not we're doing a one- or two- tail test (two-tail test is most typical).

Example

- Test to determine if the slope is significantly different from "0" (at 5% significance level)
- We want to test:
  - H_1: $\beta_1 \neq 0$
  - H_0: $\beta_1 = 0$
- If the null hypothesis is true, no linear relationship exists
- The rejection region is:

$$t < -t_{\alpha/2,v} = -t_{.025,98} \approx -1.984 \quad or \quad t > t_{\alpha/2,v} = t_{.025,98} \approx 1.984$$

- OR check the p-value.

Example

Coefficients:

Estimate Std. Error t value $Pr(>|t|)$

(Intercept)   6.27614   1.06281   5.905 7.01e-07 ***

X110m.hurdle 0.32329   0.07273   4.445 7.08e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.217 on 39 degrees of freedom

Multiple R-squared: 0.3363,   Adjusted R-squared: 0.3193

F-statistic: 19.76 on 1 and 39 DF, p-value: 7.082e-05

☐ We can see the p-value of the t-statistic is almost 0.

☐ We can say that there are overwhelming evidences to infer that a linear relation between both variables exists.

Slope confidence interval

- We can also estimate (to some level of confidence) and interval for the slope parameter, β_1 .
- Recall that your estimate for β_1 is b_1.
- The confidence interval estimator is given as:

$$b_1 \pm t_{\alpha/2} s_{b_1} \qquad s_{b_1} = \frac{s_{\varepsilon}}{\sqrt{(n-1)s_x^2}}$$

- Her $b_1 \pm t_{\alpha/2} s_{b_1} = -.0669 \pm 1.984(.00497) = -.0669 \pm .0099$

- That is, we estimate that the *slope coefficient* lies between
- $-.0768$ and $-.0570$

Coefficient of Determination

- To measure the strength of the relationship
  - We shown if a linear relationship exists, but it is also needed to measure the strength of this relationship
  - This is done by calculating the coefficient of determination

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \ or \ R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

- The coefficient of determination (R^2) is the square of the coefficient of correlation (**r**), hence $R^2 = (r)^2$

Coefficient of Determination

- If R^2 has a value of .6483. This means **64.83% of the variation of (y) is explained** by your regression model. The remaining 35.17% is unexplained, i.e. due to error.

- Unlike the value of a test statistic, the coefficient of determination **does not have a critical value** that enables us to draw conclusions.

- In general the **higher** the value of R^2, the **better** the model fits the data.

  - $R^2 = 1$: Perfect match between the line and the data points.
  - $R^2 = 0$: There are no linear relationship between x and y.

The adjusted R^2

- To avoid the increasing of the R^2 value when extra explanatory variables are added to the model.

- The adjusted R^2 can be negative

- Its value will always be less than or equal to that of R^2.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1}$$

P: total number of explanatory variables.
n: the sample size

Using the model

Intervals

Using the Regression Equation

- We could use our regression equation:
  - $y = 17.250 - .0669x$
- to predict y depending on a x value (x=40)
  - $y = 17.250 - .0669x = 17.250 - .0669(40) = 14,574$
- We call this value (14,574) a point prediction (**estimate**).
- We can estimate the value of y (point prediction) in terms of a **confidence interval**.

Intervals

- When you fit a parameter of a model, the accuracy or precision can be expressed as:
  - confidence interval
  - prediction interval
  - tolerance interval
- We assume that the data really are randomly sampled from a Gaussian distribution.

Confidence intervals

- **Confidence intervals** tell you about how well you have determined the mean. If you do this many times, and calculate a confidence interval of the mean from each sample, you'd expect **about 95** % of those intervals to **include** the true value of the **population mean.**
  - The key point is that the confidence interval tells you about the likely location of the true population parameter.

Confidence Interval: estimator for Mean of Y

- The **confidence interval** estimate for the expected value of y (Mean of Y) is used when we want to predict an interval we are pretty sure contains the true "regression line" . In this case, we are estimating the mean of y given a value of x:

$$\hat{y} \pm t_{\alpha/2, n-2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

Prediction intervals

- **Prediction intervals** tell you where you can expect to see the next data point sampled. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you'd expect that next value to lie within that prediction interval in 95% of the samples.

  - The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

  - Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So a prediction interval is always wider than a confidence interval.

Prediction Interval

- The **prediction interval** is used when we want to predict one particular value of the dependent variable, given a specific value of the independent variable:

$$\hat{y} \pm t_{\alpha/2, n-2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

$$\hat{y} = b_0 + b_1 x_g$$

- (x_g is the given value of x we're interested in)

What's the Difference?

Prediction Interval	Confidence Interval
☐ Used to estimate the value of one value of y (at given x)	☐ Used to estimate the mean value of y (at given x)

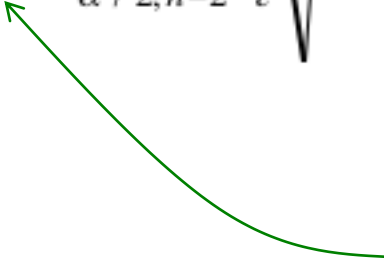$$\hat{y} \pm t_{\alpha/2, n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

$$\hat{y} \pm t_{\alpha/2, n-2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

The confidence interval estimate of the expected value of y will be narrower than the prediction interval for the same given value of x and confidence level. This is because there is less error in estimating a mean value as opposed to predicting an individual value.

Tolerance interval

- The word 'expect' used in defining of a prediction interval means there is a 50% chance that you'd see the value within the interval in more than 95% of the samples, and a 50% chance that you'd see the value within the interval in less than 95% of the samples.

- As an example doing lots of simulations, you know the true value and thus know if it is in the prediction interval or not. Hence you can then tabulate what fraction of the time the value is enclosed by the interval.

- On average the obtained value will be 95%, but it might be 92% or 98%. That means that half the time it will be less than 95% and half the time it will be more than 95%.

Tolerance interval

- If you want to be 95% sure that the interval contains 95% of the values, or 91% sure that the interval contains 99% of the values, you need the tolerance interval.

- To compute, or understand, a tolerance interval you have to specify two different percentages:
  - (i) one expresses how sure you want to be, and
  - (ii) other to expresses what fraction of the values the interval will contain.

- If you set the first value (how sure) to 50%, the tolerance interval is the prediction interval. If you set it to a higher value (say 80% or 99%) then the tolerance interval is wider.

- http://cran.r-project.org/web/packages/tolerance/tolerance.pdf

Regression diagnostics

Required Conditions

□ For these regression methods to be valid the following conditions for the error variable (ε) must be met:

- The probability distribution of ε is **normal.**
  - The **mean** of the distribution is **0**; that is, $E(\varepsilon) = 0$.
- The **standard deviation** of ε is σ_ε, which is a **constant** regardless of the value of x.
- The **value of** ε associated with any particular value of y **is independent** of ε associated with any other value of y.

Regression Diagnostics

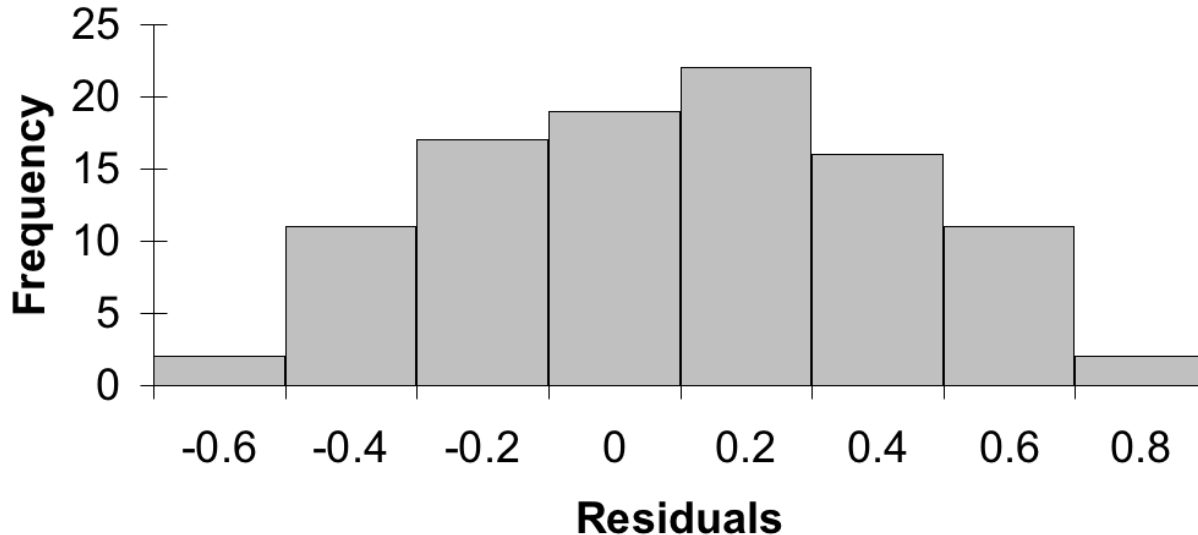- How can we diagnose violations of these conditions?
  - **Residual Analysis,** that is, examine the differences between the actual data points and those predicted by the linear equation.

Non-normality

- We can take the residuals and put them into a histogram to visually check for normality…

- …we're looking for a bell shaped histogram with the mean close to zero [our old "test for normality].

Box Cox Transformation

- A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape.
  - If your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.
- The Box Cox transformation is named after statisticians George Box and Sir David Roxbee Cox who collaborated on a 1964 paper and developed the technique.

How to apply the technique

- https://www.statisticshowto.datasciencecentral.com/box-cox-transformation/

- https://www.researchgate.net/publication/268412346_The_Box-Cox_Transformation_Technique_A_Review

- https://www.frontiersin.org/articles/10.3389/fams.2015.00012/full

Heteroscedasticity

- When the requirement of a constant variance is violated, we have a condition of heteroscedasticity.

- We desire homocedasticity.

- We can diagnose heteroscedasticity by plotting the residual against the predicted y.

- If the variance of the error variable ($\sigma_\varepsilon{}^2$) is not constant, then we have "heteroscedasticity".

Homocedasticity

□ If the variance of the error variable (σ_ε^2) is constant, then we have "homocedasticity".
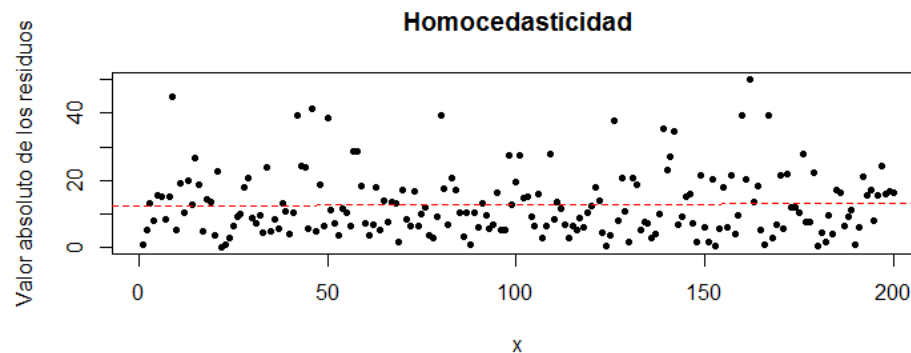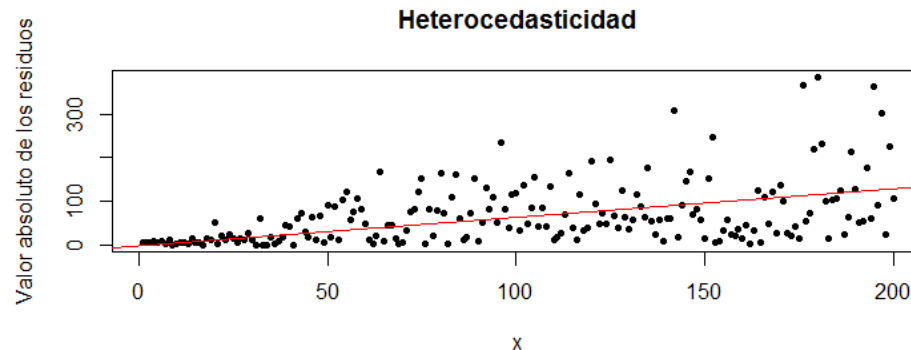
Nonindependence of the Error Variable

- When the data are time series, the errors often are correlated. Error terms that are correlated over time are said to be autocorrelated or serially correlated.

- We can often detect autocorrelation by graphing the residuals against the time periods. **If a pattern emerges,** it is likely that **the independence** requirement **is violated**.

Nonindependence of the Error Variable

- Patterns in the appearance of the residuals over time indicates that autocorrelation exists:
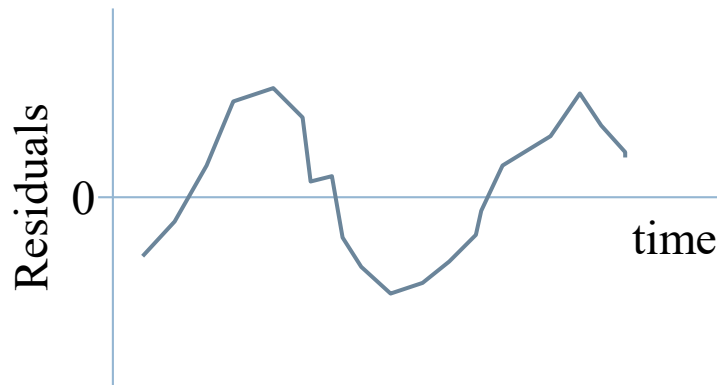
Note the runs of positive residuals, replaced by runs of negative residuals
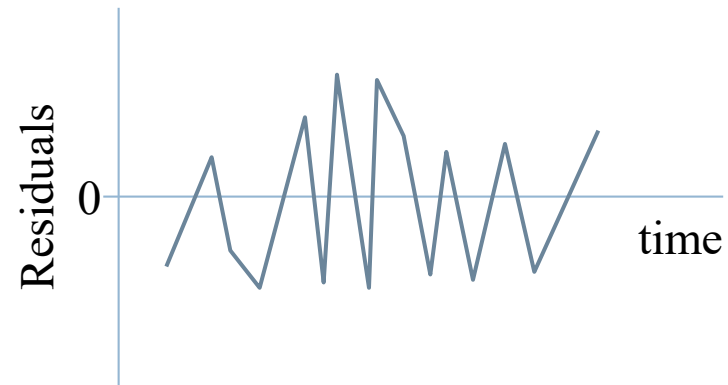
Note the oscillating behavior of the residuals around zero

Outliers

- An outlier is an observation that is unusually small or unusually large.
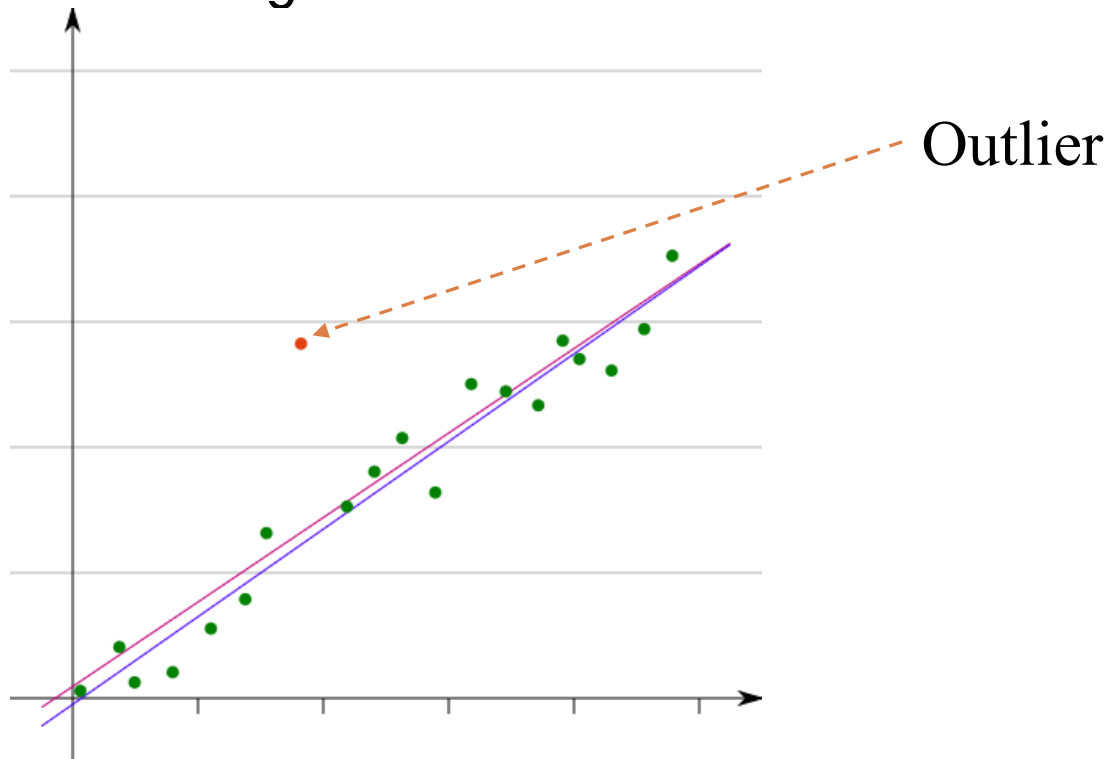
Outlier

Outliers

- Possible reasons for the existence of outliers include:
    - There was an error in recording the value
    - The point should not have been included in the sample
    - Perhaps the observation is indeed valid.
- Outliers can be easily identified from a scatter plot.
- If the absolute value of the standard residual is > 2, we suspect the point may be an outlier and investigate further.
- They need to be dealt with since they can easily influence the least squares line.

Procedure for Regression Diagnostics

- Develop a model that has a theoretical basis.
- Gather data for the two variables in the model.
- Draw the scatter diagram to determine whether a linear model appears to be appropriate. Identify possible outliers.
- Determine the regression equation.
- Calculate the residuals and check the required conditions
- Assess the model's fit.
- If the model fits the data, use the regression equation to predict a particular value of the dependent variable and/or estimate its mean.

Assumptions

- The observations within each sample must be **independent.**
  - Durbin Watson test
  - dwtest(RegModel.3, alternative = "two.sided")
- The populations from which the samples are selected must be **normal.**
  - Shapiro test
  - shapiro.test(residuals(regModel.3))
- The populations from which the samples are selected must have equal variances (**homogeneity** of variance)
  - Breusch Pagan test
  - lmtest::bptest(Regmodel.3)

Example

Do it by hand

Linear regression

x	y
5	5
10	2
4	8
8	3
2	8
7	5
9	5
6	7
1	10
12	3

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

Solution

x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	S_{xy}	S_x^2
5	5	-1,4	-0,6	0,84	1,96
10	2	3,6	-3,6	-12,96	12,96
4	8	-2,4	2,4	-5,76	5,76
8	3	1,6	-2,6	-4,16	2,56
2	8	-4,4	2,4	-10,56	19,36
7	5	0,6	-0,6	-0,36	0,36
9	5	2,6	-0,6	-1,56	6,76
6	7	-0,4	1,4	-0,56	0,16
1	10	-5,4	4,4	-23,76	29,16
12	3	5,6	-2,6	-14,56	31,36
64	56			**-73,4**	**110,4**
$\bar{x}=$**6,4**	$\bar{y}=$**5,6**				
				b1=	**-0,66486**
				b0=	**9,855072**

Calculate the parameters

- $S_{xy} = -73.4/n-1$
- $S_x^2 = 110.4/n-1$
- $b_1 = -73.4/110.4 = -0.665$
- $b_0 = 5.6 - (-0.665(6.4)) = 9.856$

Estimate the prediction interval

- Remember

$$\hat{y} \pm t_{\alpha/2, n-2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

- Calculate the value for 11.

- Interpret this value

- $t_{.025,8} = 2.306$

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right)$$

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-2}}$$

Estimate the prediction interval

- Calculate the value for 11.
- First we calculate the predicted value
  - y=9.856-0.665(11)=2.541
- Second calculate the value for a $t_{.025,8} = 2.306$
- Third, proceed with the formula
- $(a,b) =$

- $= 2.541 \pm 2.306(1.204)\sqrt{1 + \frac{1}{10} + \frac{(11-6.4)^2}{110.40}}$

- $=2.541 \pm 3.155$
- (-0.614,5.696)

Estimate the interval for the slope

- The values are:

- $t_{.025,8} = 2.306$

- $(a, b) =$

- $-0.665 \pm 2.306 \dfrac{1.204}{\sqrt{110.4}} = (-0.929, -0.401)$

$$b_1 \pm t_{\alpha/2} s_{b_1} \quad \nu = n - 2 \qquad s_{b_1} = \dfrac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$