Week 13: A Primer on Linear Regression

MATH697

Sahir Bhatnagar

November 28, 2017

McGill University

Motivating Example: Blood pressure

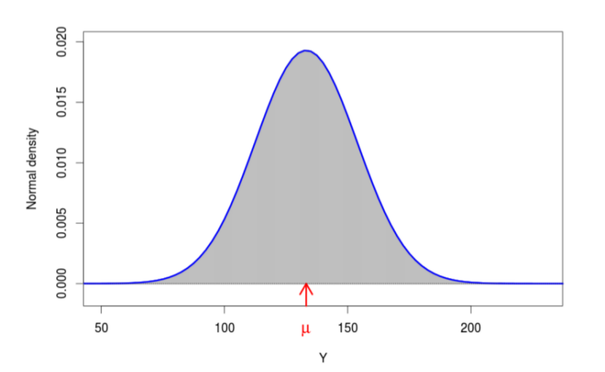
Repeated measurements of systolic blood pressure

- We have two consecutive systolic blood pressure measurements for 1999 individuals. Let X denote the first measurement and Y the second one.
- The observed data consist of pairs (x_i, y_i) for $i = 1, \ldots, 1999$.
- The arithmetic means (standard deviations) are $\bar{x} = 133.2\ (20.7)$ and $\bar{y} = 132.3\ (20.1)$.
- Consider the task of **predicting** the systolic BP at the second measurement, given the first measurement.
- In other words, we are interested in the **expected systolic BP at the second measurement**, given the first measurement X. We denote this quantity as μ_X.

[1] This motivating example is from Olli Saarela at UofT

What do we mean by the expected value?

- The expected value μ is the centre of gravity of the theoretical distribution of Y:



- If placed on the arrow head at the expected value, the distribution would be balanced, i.e. not tipping over to either side.

Conditional expectation

- The expected value of Y conditional on X, μ_X, is the centre of gravity of the theoretical distribution of Y in a subpopulation where everyone has the initial systolic BP measured to be exactly X.
- μ_X is our prediction for Y based on the measured value of X.
- This prediction is derived from a regression equation, say,

$$\mu_X = \beta_0 + \beta_1 X$$

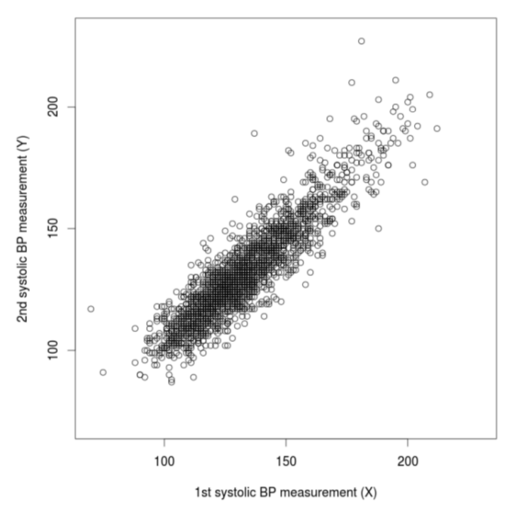
- Complemented with a statistical distribution, say,

$$Y_X \sim N(\beta_0 + \beta_1 X, \sigma^2),$$

  we have a regression model that can be fitted to the data to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$.

The data

- The empirical bivariate distribution may presented as a *scatter plot*:



Results

```
> model <- lm(y ~ x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-30.496  -5.277  -0.247   4.967  53.327

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.265613   1.245484   12.26   <2e-16 ***
x            0.878883   0.009243   95.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.54 on 1997 degrees of freedom
Multiple R-squared:  0.8191,    Adjusted R-squared:  0.819
F-statistic:  9042 on 1 and 1997 DF,  p-value: < 2.2e-16
```

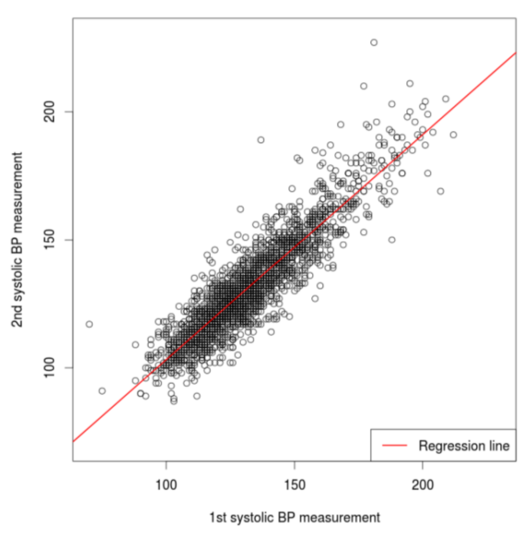
Calculating the predictions

- From the regression estimates, we may now calculate the predictions as

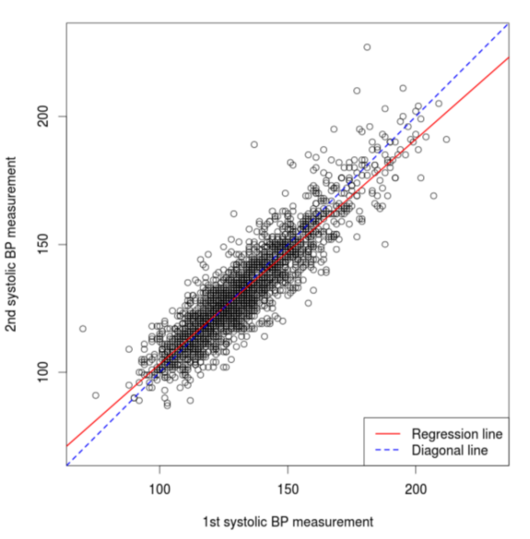
$$\begin{aligned}\hat{\mu}_X &= \hat{\beta}_0 + \hat{\beta}_1 X \\ &= 15.266 + 0.879X.\end{aligned}$$

- The *fitted values* $\hat{\mu}_X$ may be calculated at each possible value of X.
- This results in a *regression line* with an intercept of 15.266 and a slope of 0.879.

Regression line



Add a diagonal line (intercept = 0, slope = 1)



Interpretation

- With a one mmHg increase (decrease) in the first measurement, the prediction of the second is increased (decreased) by 0.879 mmHg.

Interpretation

- With a one mmHg increase (decrease) in the first measurement, the prediction of the second is increased (decreased) by 0.879 mmHg.
- With high initial systolic BP measurement *X*, the predicted value from the regression model is **smaller** than *X*.
- Conversely, with a low initial systolic BP measurement *X*, the predicted value from the regression model is **higher** than *X*.

Interpretation

- With a one mmHg increase (decrease) in the first measurement, the prediction of the second is increased (decreased) by 0.879 mmHg.
- With high initial systolic BP measurement X, the predicted value from the regression model is **smaller** than X.
- Conversely, with a low initial systolic BP measurement X, the predicted value from the regression model is **higher** than X.
- For instance, if $X = 175$, $\hat{\mu}_{175} = 15.266 + 0.879 \times 175 \approx 169.1$.
- If $X = 100$, $\hat{\mu}_{100} = 15.266 + 0.879 \times 100 \approx 103.2$.
- How can this be?

Interpretation

- With a one mmHg increase (decrease) in the first measurement, the prediction of the second is increased (decreased) by 0.879 mmHg.
- With high initial systolic BP measurement X, the predicted value from the regression model is **smaller** than X.
- Conversely, with a low initial systolic BP measurement X, the predicted value from the regression model is **higher** than X.
- For instance, if $X = 175$, $\hat{\mu}_{175} = 15.266 + 0.879 \times 175 \approx 169.1$.
- If $X = 100$, $\hat{\mu}_{100} = 15.266 + 0.879 \times 100 \approx 103.2$.
- How can this be?
- Could this be explained by the fact that the second BP value is on average slightly lower than the first? (Why?)

Interpretation (continued)

- First, is the regression coefficient significantly different from one?
- The null hypothesis is $\beta = 1$, and the corresponding z-statistic and p-value are $Z = \frac{0.879-1}{0.00924} = -13.1$ and

```
2 * pnorm(-13.1)
[1] 3.292577e-39
```

- This is clearly a significant difference from one.
- So what is the explanation?

Interpretation (continued)

- First, is the regression coefficient significantly different from one?
- The null hypothesis is $\beta = 1$, and the corresponding z-statistic and p-value are $Z = \frac{0.879-1}{0.00924} = -13.1$ and

  ```
  2 * pnorm(-13.1)
  [1] 3.292577e-39
  ```

- This is clearly a significant difference from one.
- So what is the explanation?
- Recall that measured blood pressure values contain **measurement error**, which is why the two measurements are not exactly the same, even if the underlying **true** blood pressure would remain constant between the measurements.

Measurement error

- A very high initial blood pressure measurement is likely a **combination of two factors**, a high underlying true blood pressure, and positive measurement error.
- Assuming that there is no systematic measurement error to either direction, there is no reason why the error in the second measurement would be equal in magnitude and sign.

Measurement error

- A very high initial blood pressure measurement is likely a **combination of two factors**, a high underlying true blood pressure, and positive measurement error.
- Assuming that there is no systematic measurement error to either direction, there is no reason why the error in the second measurement would be equal in magnitude and sign.
- Thus, in terms of obtaining the best prediction, it makes sense to slightly scale down (up) the value of the first measurement, if it is extreme to either direction.

Measurement error

- A very high initial blood pressure measurement is likely a **combination of two factors**, a high underlying true blood pressure, and positive measurement error.
- Assuming that there is no systematic measurement error to either direction, there is no reason why the error in the second measurement would be equal in magnitude and sign.
- Thus, in terms of obtaining the best prediction, it makes sense to slightly scale down (up) the value of the first measurement, if it is extreme to either direction.
- This is exactly what fitting a linear regression model does for us.

Regression towards the mean

- This phenomenon was discovered by Sir Francis Galton in late 19th century.
- He was studying how the height of the offspring depends on the height of the parents, and noted that the children of tall parents tend to be shorter than their parents, by a constant fraction, the regression coefficient.

Regression towards the mean

- This phenomenon was discovered by Sir Francis Galton in late 19th century.
- He was studying how the height of the offspring depends on the height of the parents, and noted that the children of tall parents tend to be shorter than their parents, by a constant fraction, the regression coefficient.
- The explanation here is not measurement error as such; rather, the extreme stature of a given parent is caused by a number of pairs of recessive alleles, with two copies of any of these alleles (homozygous genotypes) contributing (say) additively to the height phenotype.
- However, since the two parents do not necessarily share the same homozygous genotypes (even if they have equal amount of them), the offspring is unlikely to inherit as many stature related homozygous

Interpreting the regression coefficient

- Compare the predictions for two individuals who differ in their initial systolic BP by only 1 mmHg.
- Now

$$\mu_{X+1} - \mu_X = \beta_0 + \beta_1 X - [\beta_0 + \beta_1 (X - 1)] = \beta_1$$

- The regression coefficient is the **change in the expected outcome** when the value of the predictor changes one unit.

Interpreting the regression coefficient

- Compare the predictions for two individuals who differ in their initial systolic BP by only 1 mmHg.
- Now

$$\mu_{X+1} - \mu_X = \beta_0 + \beta_1 X - [\beta_0 + \beta_1(X-1)] = \beta_1$$

- The regression coefficient is the **change in the expected outcome** when the value of the predictor changes one unit.
- Suppose that instead of the initial blood pressure measurement, the predictor is binary, say, current use of antihypertensive medication (1 indicating current use and 0 non-use). Now

$$\mu_1 - \mu_0 = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

Interpreting the regression coefficient

- Compare the predictions for two individuals who differ in their initial systolic BP by only 1 mmHg.
- Now

$$\mu_{X+1} - \mu_X = \beta_0 + \beta_1 X - [\beta_0 + \beta_1(X - 1)] = \beta_1$$

- The regression coefficient is the **change in the expected outcome** when the value of the predictor changes one unit.
- Suppose that instead of the initial blood pressure measurement, the predictor is binary, say, current use of antihypertensive medication (1 indicating current use and 0 non-use). Now

$$\mu_1 - \mu_0 = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

- Let us fit such a model, also adjusting for age and sex.

Results

```
> model <- lm(y ~ meds + age + sex)
> summary(model)

Call:
lm(formula = y ~ meds + age + sex)

Residuals:
    Min      1Q  Median      3Q     Max
-51.195 -13.522  -1.939  10.559  82.439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.25289    3.74044  24.396  < 2e-16 ***
meds        12.03391    1.60958   7.476 1.14e-13 ***
age          0.75044    0.07403  10.137  < 2e-16 ***
sex          4.66675    0.85724   5.444 5.86e-08 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 19.11 on 1995 degrees of freedom
Multiple R-squared:  0.09514,   Adjusted R-squared:  0.09378
```

Calculating predictions from the regression equation

- Any prediction may now be calculated from the estimated regression equation as

$$\hat{\mu} = 91.25 + 12.03 \times \mathrm{meds} + 0.75 \times \mathrm{age} + 4.67 \times \mathrm{sex}.$$

Calculating predictions from the regression equation

- Any prediction may now be calculated from the estimated regression equation as

$$\hat{\mu} = 91.25 + 12.03 \times \text{meds} + 0.75 \times \text{age} + 4.67 \times \text{sex}.$$

- For example, suppose we want to compare the expected systolic BP of a 55-year old male on antihypertensive medication to a 55-year old male not on medication.
- We get

$$91.25 + 12.03 \times 1 + 0.75 \times 55 + 4.67 \times 1 = 149.2$$

  and

$$91.25 + 12.03 \times 0 + 0.75 \times 55 + 4.67 \times 1 = 137.17.$$

- The difference is 149.2 - 137.17 = 12.03, the regression coefficient.

Correlation coefficient

- As we have seen, the interpretation of the regression coefficient depends on the measurement scales of the predictor and outcome variables.
- A close relative of the regression coefficient is the correlation coefficient, where the measurement scales have been eliminated by **rescaling** by the standard deviations of the two variables.
- The theoretical correlation coefficient is defined as

$$\rho = \beta_1 \frac{\sigma_X}{\sigma_Y},$$

  and its estimator by

$$r = \hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}}.$$

Correlation coefficient

- As we have seen, the interpretation of the regression coefficient depends on the measurement scales of the predictor and outcome variables.
- A close relative of the regression coefficient is the correlation coefficient, where the measurement scales have been eliminated by **rescaling** by the standard deviations of the two variables.
- The theoretical correlation coefficient is defined as

$$\rho = \beta_1 \frac{\sigma_X}{\sigma_Y},$$

and its estimator by

$$r = \hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}}.$$

Correlation coefficient (continued)

- Conversely, we may calculate the regression coefficient estimate from the relationship $\hat{\beta}_1 = r\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, and the intercept term from the relationship $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- Correlation coefficient takes values in the $[-1, 1]$ interval, with 1 (-1) meaning perfect positive (negative) correlation.

Correlation coefficient (continued)

- Conversely, we may calculate the regression coefficient estimate from the relationship $\hat{\beta}_1 = r\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, and the intercept term from the relationship $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- Correlation coefficient takes values in the $[-1, 1]$ interval, with 1 (-1) meaning perfect positive (negative) correlation.
- An important correlation coefficient related to linear regression is $R = \mathrm{cor}(Y, \hat{\mu}_X)$, that is, the **correlation between the outcome variable and its prediction**. This is a measure of the model fit; the more informative the predictors are of the outcome, the closer this correlation is to one.

Correlation coefficient (continued)

- Since this correlation does not take negative values, the usually reported quantity is its square, R^2.
- This measures what proportion of the variation of the outcome variable is explained by the regression equation.

Explained and unexplained variation

- The R^2 is calculated in the R summary output.
- The model with the first systolic BP measurement as the only predictor gives $R^2 = 0.8191$, while the model with antihypertensive drug use, age and sex gives $R^2 = 0.09514$.

Explained and unexplained variation

- The R^2 is calculated in the R summary output.
- The model with the first systolic BP measurement as the only predictor gives $R^2 = 0.8191$, while the model with antihypertensive drug use, age and sex gives $R^2 = 0.09514$.
- Recall that the regression equation explains only part of the variation of the outcome; the remaining unexplained variation is characterized by the assumed statistical distribution.

Explained and unexplained variation

- The R^2 is calculated in the R summary output.
- The model with the first systolic BP measurement as the only predictor gives $R^2 = 0.8191$, while the model with antihypertensive drug use, age and sex gives $R^2 = 0.09514$.
- Recall that the regression equation explains only part of the variation of the outcome; the remaining unexplained variation is characterized by the assumed statistical distribution.
- The statistical distribution is now $Y_X \sim N(\mu_X, \sigma^2)$, or equivalently, $Y_X - \mu_X \sim N(0, \sigma^2)$.

Explained and unexplained variation

- The R^2 is calculated in the R summary output.
- The model with the first systolic BP measurement as the only predictor gives $R^2 = 0.8191$, while the model with antihypertensive drug use, age and sex gives $R^2 = 0.09514$.
- Recall that the regression equation explains only part of the variation of the outcome; the remaining unexplained variation is characterized by the assumed statistical distribution.
- The statistical distribution is now $Y_X \sim N(\mu_X, \sigma^2)$, or equivalently, $Y_X - \mu_X \sim N(0, \sigma^2)$.
- The quantity $\varepsilon = Y_X - \mu_X$ is the prediction error, and its statistical distribution characterizes the unexplained variation.

Explained and unexplained variation

- The R^2 is calculated in the R summary output.
- The model with the first systolic BP measurement as the only predictor gives $R^2 = 0.8191$, while the model with antihypertensive drug use, age and sex gives $R^2 = 0.09514$.
- Recall that the regression equation explains only part of the variation of the outcome; the remaining unexplained variation is characterized by the assumed statistical distribution.
- The statistical distribution is now $Y_X \sim N(\mu_X, \sigma^2)$, or equivalently, $Y_X - \mu_X \sim N(0, \sigma^2)$.
- The quantity $\varepsilon = Y_X - \mu_X$ is the prediction error, and its statistical distribution characterizes the unexplained variation.
- The empirical version of this quantity, $\hat{\varepsilon}_i = y_i - \hat{\mu}_{x_i}$, is known as the *residual*, and can be used for checking the above normality assumption.

More on residuals

- The residual standard error in the R summary output is an estimate of the standard deviation σ in $\varepsilon \sim N(0, \sigma^2)$.
- In the ongoing example, we can visually compare the residuals to $N(0, 8.54^2)$-distribution by for example comparing the histogram to the normal curve.

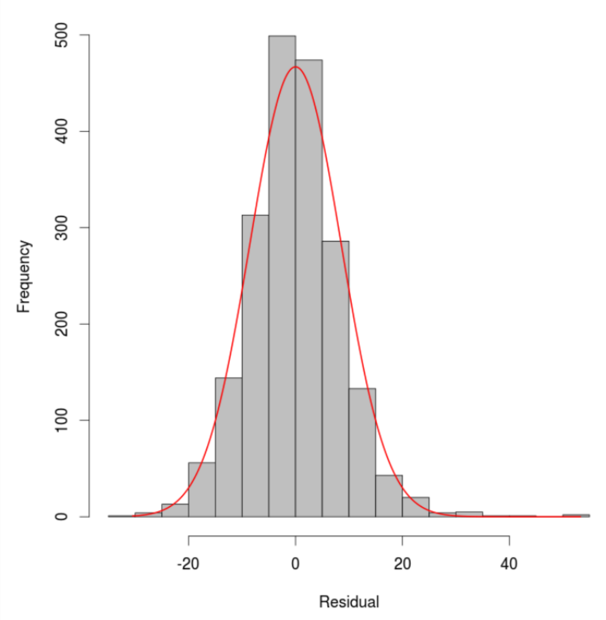
More on residuals

- The residual standard error in the R summary output is an estimate of the standard deviation σ in $\varepsilon \sim N(0, \sigma^2)$.
- In the ongoing example, we can visually compare the residuals to $N(0, 8.54^2)$-distribution by for example comparing the histogram to the normal curve.
- Another alternative is the *quantile-quantile (QQ) plot*, which should resemble a straight line when the empirical distribution is close to normal.
- The QQ-plot presentation is much more sensitive to deviations from normality in the tails of the distribution.

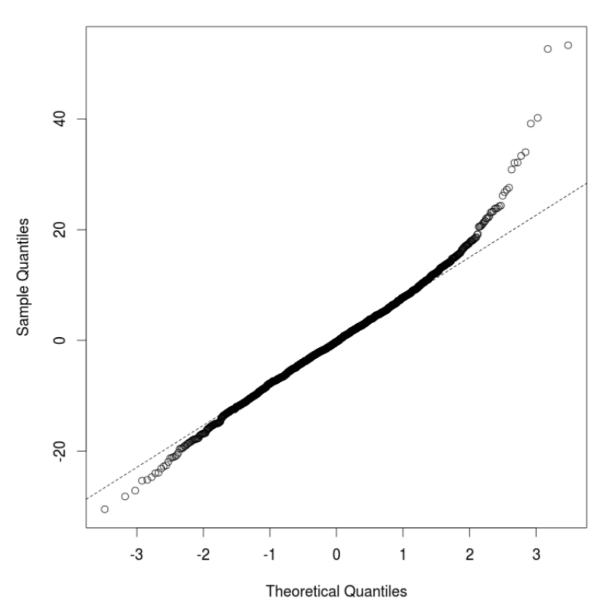
More on residuals (continued)

- Normality of the residuals is not relevant to the interpretation of the regression coefficients. The standard errors, however, are based on the normality assumption, so large deviations might require (say, log-) transformation of the outcome variable to restore normality.
- Note that such a transformation also **changes the interpretation of the regression coefficient**.

Histogram with a normal curve



QQ-plot



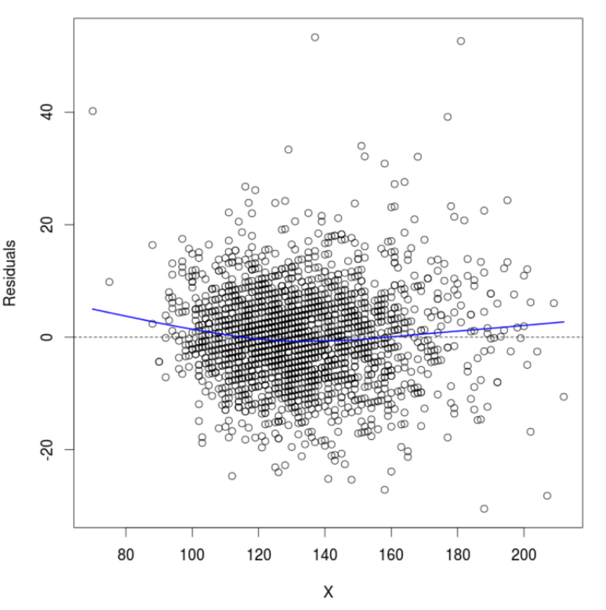
Further residual checks

- There is another, perhaps less explicit assumption involved in the specification $\varepsilon \sim N(0, \sigma^2)$.
- This assumes that the residual standard deviation σ is constant across different values of X.

Further residual checks

- There is another, perhaps less explicit assumption involved in the specification $\varepsilon \sim N(0, \sigma^2)$.
- This assumes that the residual standard deviation σ is constant across different values of X.
- This property is known as **homoschedasticity**, and if this assumption is not satisfied, the residuals exhibit **heteroschedasticity**.
- This can be checked simply by plotting the residuals against the X-variable.
- The plot should not show any systematic characteristics, for instance, a funnel-like shape.
- Visual inspection can be aided by adding a some kind of moving average, for example a LOWESS-curve.

Residual plot with a LOWESS-curve



Reality check

- The objective of a study is not to **study** something.
- In the same vein, **modeling**, including model selection, and checking, or testing, the **correctness** of a model, should not be an end in itself.
- By definition, there is no such thing as a correct model; a model is a simplification of reality, not the reality itself.

Reality check

- The objective of a study is not to **study** something.
- In the same vein, **modeling**, including model selection, and checking, or testing, the **correctness** of a model, should not be an end in itself.
- By definition, there is no such thing as a correct model; a model is a simplification of reality, not the reality itself.
- Hence, a model need not capture all features of reality; if it could, it would no longer be a model.

Reality check

- The objective of a study is not to **study** something.
- In the same vein, **modeling**, including model selection, and checking, or testing, the **correctness** of a model, should not be an end in itself.
- By definition, there is no such thing as a correct model; a model is a simplification of reality, not the reality itself.
- Hence, a model need not capture all features of reality; if it could, it would no longer be a model.
- A good model is a model that is useful by serving some purpose.
- The purpose is to advance the practice of (community) medicine, usually by producing (diagnostic, prognostic or etiognostic) evidence.
- Some models may be better than others, depending on the chosen criterion for **better**.

The Method of Maximum Likelihood for Simple Linear Regression

Simple Linear Regression

Example 1 (Simple Linear Regression)

Suppose that y_i, $i = 1, \ldots, n$ are n independent random variables, each corresponds to a known explanatory variable x_i and has the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for some parameters β_0, β_1. The random noise $\boldsymbol{\epsilon} \sim N\left(0, \sigma^2\right)$ is independent of $\mathbf{x}$ and independent across observations. Find the MLEs for the parameters $\Theta = \left(\beta_0, \beta_1, \sigma^2\right)$.

Properties of Simple Linear Regression Parameters

- The MLEs for simple linear regression can be represented as **linear combinations** of the observations y_i:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \sum c_i \bar{x} y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - c_i \bar{x}\right) y_i$$

where $c_i = (x_i - \bar{x})/S_{xx}$

Properties of Simple Linear Regression Parameters

- The MLEs for $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are **unbiased estimators** of the model parameters β_1 and β_0. To show this consider

$$E(\widehat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} ci(\beta_0 + \beta_1 x_i) = \beta_1$$

$$E(\widehat{\beta}_0) = E\left(\sum_{i=1}^{n} \left(\frac{1}{n} - c_i \overline{x}\right) y_i\right) = \beta_0$$

since $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$

Properties of Simple Linear Regression Parameters

- It can be shown that the MLE for $\widehat{\sigma}^2 = \left(1/n\right) \sum \hat{\epsilon}_i^2$ is a **biased** estimator for σ^2. We instead use the **unbiased** estimator given by

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

Properties of Simple Linear Regression Parameters

- The variance of $\widehat{\beta}_1$ is found as

$$Var(\widehat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 Var(y_i) = \frac{\sigma^2}{S_{xx}} = \frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{xx}}$$

  since the observations y_i are uncorrelated, and so the variance of the sum is just the sum of the variances, and we have plugged in the unbiased estimator for σ^2.

- The variance of $\widehat{\beta}_0$ is found as

$$Var(\widehat{\beta}_0) = Var\left(\sum_{i=1}^{n}\left(\frac{1}{n} - c_i\bar{x}\right) y_i\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Recap

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$Var(\widehat{\beta}_1) = \frac{\widehat{\sigma}^2}{S_{xx}}, \quad S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

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

$$Var(\widehat{\beta}_0) = \widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$$

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightarrow \text{unbiased estimator of } \sigma^2$$

Simple Linear Regression

Example 2 (Simple Linear Regression Women dataset)

Use the `data(women)` included in the `datasets` package. Let x : weight and y : height. Define the relationship

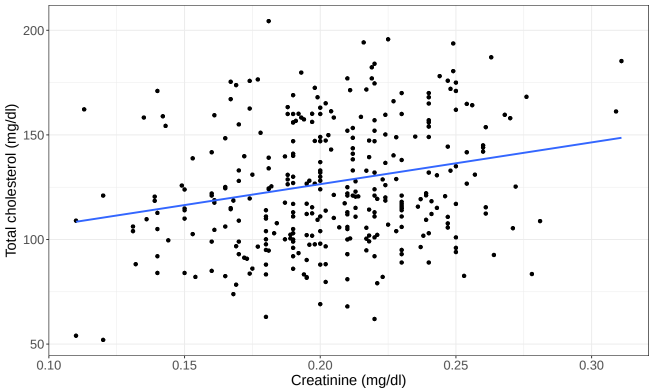
$$\text{height} = \beta_0 + \beta_1 \cdot \text{weight} + \varepsilon$$

Find the maximum likelihood estimates and variances for β_0 and β_1 using the formulas on the previous slide. Compare your answer with the `lm` function in R.

```r
data(women) # from the datasets package
fit <- lm(height ~ weight, data = women); summary(fit)
```

Least Squares: An Alternative to MLE

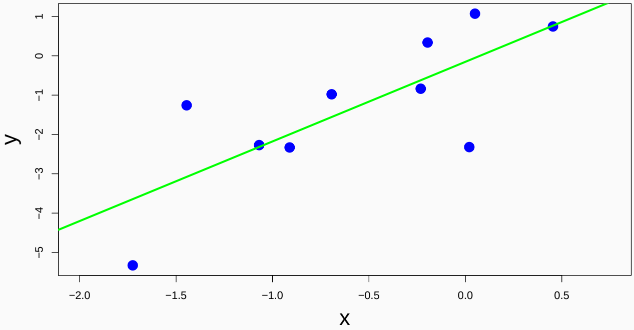
Scatterplot with best fit line



Want to find a "best fit" line to the data. The slope of the line is usually the parameter of interest.

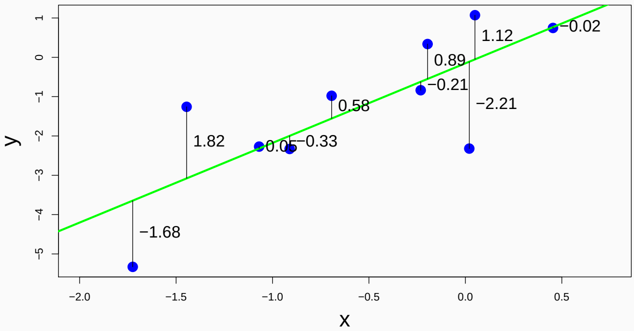
[1] Slides from Kevin McGregor at McGill

How to find best fit line?



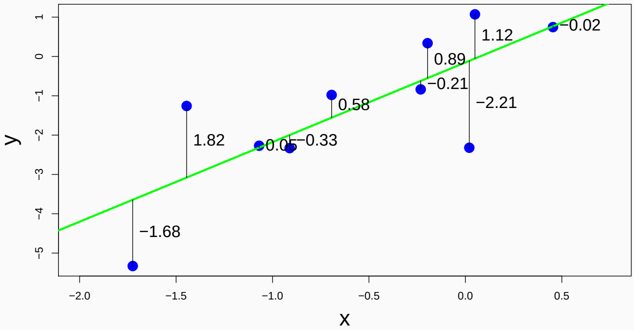
How do we find a best fit line? Depends on how we define “best” fit.

Residuals



Recall the residuals (vertical distance between the true point and the fitted line). Assume *x* values are fixed (no measurement error).

Squared residuals



Consider the sum squared residuals: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the value falling on the line directly above or below y_i.

Least-squares estimator

- Turns out there is a nice solution to this problem, called the least-squares estimator.
- Assume that the regression line takes on the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_1 estimates the slope of the line... or the effect of the *predictor* variable x on the *outcome* variable y.
- We also assume that each point has an error term ϵ_i which follows a normal distribution with mean 0 and standard deviation σ.
    - Contains additional variation in y not explained by the predictor variable x.

Least-squares Objective Function

The least squares line minimizes the sum of squared vertical distances from the points to the line $y = \beta_0 + \beta_1 x$. We choose the slope β_1 and intercept β_0 of the straight line to minimize the **residual sum of squares** (or equivalently the **error sum of squares**)

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

The least squares estimator $\hat{\Theta}^{LS} = (\hat{\beta}_0, \hat{\beta}_1)$ is given by

$$\hat{\Theta}^{LS} = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Least-squares Objective Function

The least squares line minimizes the sum of squared vertical distances from the points to the line $y = \beta_0 + \beta_1 x$. We choose the slope β_1 and intercept β_0 of the straight line to minimize the **residual sum of squares** (or equivalently the **error sum of squares**)

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

The least squares estimator $\hat{\Theta}^{LS} = (\hat{\beta}_0, \hat{\beta}_1)$ is given by

$$\hat{\Theta}^{LS} = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Recall that the maximum likelihood estimator $\hat{\Theta}^{MLE} = (\hat{\beta}_0, \hat{\beta}_1)$ is given by

$$\hat{\Theta}^{MLE} = \arg\max_{\beta_0, \beta_1} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

Least-squares estimator

- The least squares estimator for β_1 is then:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

- The least squares estimator for β_0 is:

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

Fitted Values and Residuals

- From these LS estimates, we can calculate a **fitted value** for each individual:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

- Furthermore we can calculate the **residuals** which is the difference between the observed (y_i) and fitted value ($\hat{y}_i$):

$$\hat{\epsilon}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

Least-squares estimator of the variance

- Recall the **residual sum of squares** (or equivalently the **error sum of squares**)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

- The least squares estimator of σ^2 is given by

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Least-squares estimator of the variance

- Recall the **residual sum of squares** (or equivalently the **error sum of squares**)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

- The least squares estimator of σ^2 is given by

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

- Computation of SSE is tedious because both the **fitted values** and **residuals** must first be calculated. The following formula does not require these quantities:

$$SSE = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

Total Sum of Squares

- The error sum of squares *SSE* can be interpreted as a measure of how much variation in *y* is left unexplained by the model, i.e., how much **cannot be attributed to a linear relationship**

- A quantitative measure of the total amount of variation in observed *y* values is given by the **total sum of squares**:

$$SST = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} y_i \right)^2 / n$$

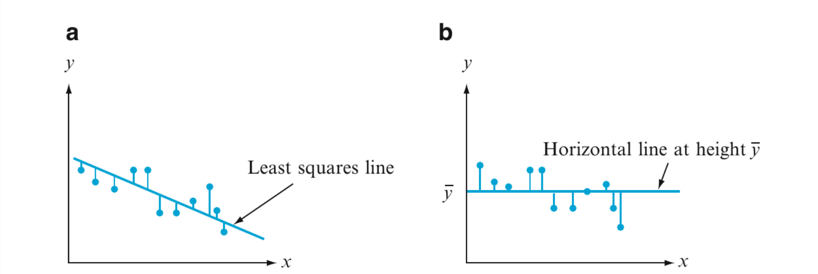


Figure 12.13 Sums of squares illustrated: (a) SSE = sum of squared deviations about the least squares line; (b) SST = sum of squared deviations about the horizontal line

Coefficient of Determination

- The ratio SSE/SST is the proportion of total variation that **cannot be explained by the simple linear regression model**
- Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{SSE}{SST}$$

  is the proportion of observed y variation **explained by the model**. This number is always between 0 and 1.

Coefficient of Determination

- The ratio SSE/SST is the proportion of total variation that **cannot be explained by the simple linear regression model**
- Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{SSE}{SST}$$

  is the proportion of observed y variation **explained by the model**. This number is always between 0 and 1.

- R^2 is the proportion by which the error sum of squares is reduced by the regression line compared to the horizontal line. For example, if $SST = 20$ and $SSE = 2$, then $R^2 = 1 - 2/20 = 0.9$. So the regression reduces the error sum of squares by 90%

Inference About the Regression Coefficient: test statistic

- The assumptions of the simple linear regression model imply that the standardized variable t_0 follows a t distribution with $n-2$ degrees of freedom (we omit the proof):

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{(n-2)}$$

where $SE(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)}$, $Var(\widehat{\beta}_1) = \widehat{\sigma}^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$ and $\widehat{\sigma}^2 = SSE/(n-2) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-2)$

Inference About the Regression Coefficient: Confidence Interval

- A $100(1-\alpha)\%$ Confidence Interval (CI) for the slope β_1 of the true regression line is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

where α is a pre-determined level of significance

Inference About the Regression Coefficient: Confidence Interval

- A $100(1-\alpha)\%$ Confidence Interval (CI) for the slope β_1 of the true regression line is given by

  $$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

  where α is a pre-determined level of significance
- The above equation implies that the following probability statement:

$$P\left(\hat{\beta}_1 - t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)\right) = 1-\alpha$$

Inference About the Regression Coefficient: Test Statistic

- Can test β_1 to check for significant linear relationship between the two variables:

  $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$

- The test statistic is given by

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{(n-2)}$$

- Testing at significance level α: compare t_0 to the critical value $t_{\alpha/2,n-2}$
  - If either $t_0 \geq t_{\alpha/2,n-2}$ or $t_0 \leq -t_{\alpha/2,n-2}$, then reject H_0.
  - Otherwise, do not reject H_0.
- p-value for this **two-sided** test is given by $2 * P(|t_0| \geq t_{\alpha,n-2})$

Least-Squares Example

Example 3 (Least-Squares)

A study was conducted to determine whether a linear relationship exists between the breaking strength y of wooden beams and the specific gravity x of the wood. Ten randomly selected beams of the same cross-sectional dimensions were stressed until they broke. Summary statistics of the breaking strengths $y_1, \ldots, y_{10}$ and the corresponding density of the wood $x_1, \ldots, x_{10}$ are $\bar{x} = 0.4951$, $\bar{y} = 11.876$, $\sum_{i=1}^{10} x_i^2 = 2.489$, $\sum_{i=1}^{10} y_i^2 = 1415.704$, $\sum_{i=1}^{10} x_i y_i = 59.207$.

(a) Fit the model $Y = \beta_0 + \beta_1 x + \varepsilon$ by least squares, and write the expression of the regression line and interpret the slope parameter.

(b) Test the null hypothesis at $\alpha = 0.05$ of whether there is a linear relationship between gravity of wood and breaking strenght. Give the null and alternative hypothesis, the test statistic, the rejection region or the p-value and conclude

(c) Give a 95% confidence interval for the slope parameter β_1

(d) Compute the R^2 and comment on the quality of the fit.

Least-Squares Example

```r
n = 10 ; xbar = 0.4951 ; ybar = 11.876
sx = n * xbar ; sy = n * ybar ;
sx2 = 2.489 ; sy2 = 1415.704 ; sxy = 59.207
(b1hat <- (sxy - sx * sy / n) / (sx2 - sx^2 / n))
```

```
## [1] 10.82958
```

```r
(b0hat <- ybar - b1hat * xbar)
```

```
## [1] 6.514273
```

```r
(SSE <- sy2 - b0hat * sy - b1hat * sxy)
```

```
## [1] 0.8817633
```

```r
(sigma2hat <- SSE / (n-2))
```

```
## [1] 0.1102204
```

```r
(seb1hat <- sqrt(sigma2hat / (sx2 - sx^2 / 10)))
```

Least-Squares Example (continued)

```
(SST <- sy2 - sy^2 / n)
```

```
## [1] 5.31024
```

```
(R2 <- 1-SSE / SST)
```

```
## [1] 0.8339504
```

```
(TestStat <- b1hat / seb1hat)
```

```
## [1] 6.338641
```

```
qt(c(0.025, 0.975), 8)
```

```
## [1] -2.306004  2.306004
```

```
(ci <- b1hat  + qt(c(0.025, 0.975), 8) * seb1hat)
```

```
## [1]  6.88977 14.76940
```

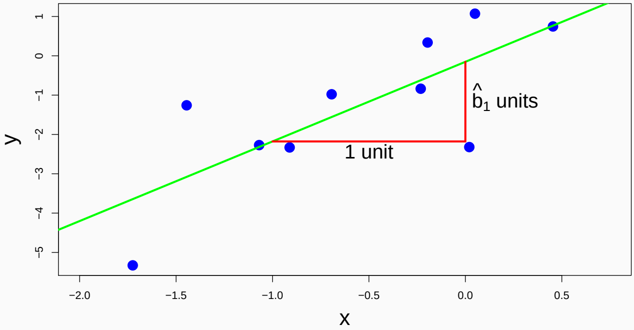
Least-squares estimator

- b_1 is usually the parameter of interest.
- Consider two individuals whose x values differ by exactly one unit (say, $x_1 = 5$ and $x_2 = 6$).
- Taking the difference of the two fitted values for these individuals gives:

$$\begin{aligned}
\hat{y}_2 - \hat{y}_1 &= (\hat{b}_0 + \hat{b}_1 x_2) - (\hat{b}_0 + \hat{b}_1 x_1) \\
&= (\hat{b}_0 + \hat{b}_1 \cdot 6) - (\hat{b}_0 + \hat{b}_1 \cdot 5) \\
&= \hat{b}_0 - \hat{b}_0 + \hat{b}_1(6 - 5) \\
&= \hat{b}_1
\end{aligned}$$

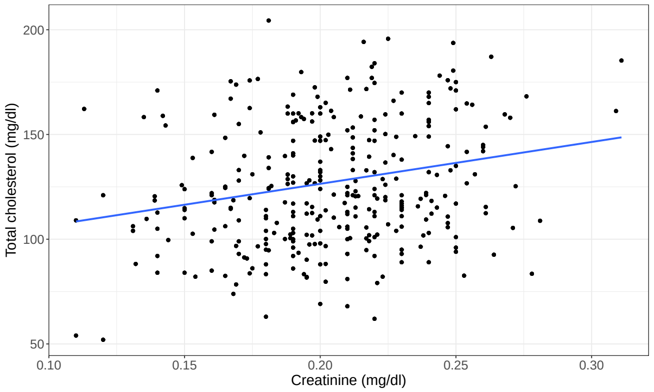
- The value $\hat{b}_1$ is the estimated change in y associated with an increase in *one unit* of x.

Coefficient visualization



Visual interpretation of $\hat{b}_1$.

IMPC data example



Example: $\hat{b}_1 = 199.74$. Means increase in 0.1 mg/dl of creatinine corresponds to an *average* increase in 19.974 mg/dl of total cholesterol.

IMPC data example significance

- In our example $\hat{b}_1 = 199.74$, $SE(\hat{b}_1) = 21.69$. Sample size is $n = 1471$. Test at level $\alpha = 0.05$.
- Calculating the t-statistic gives:

$$
\begin{aligned}
t &= \frac{\hat{b}_1}{SE(\hat{b}_1)} \\
&= \frac{199.74}{21.69} \\
&= 9.21
\end{aligned}
$$

- $t_{0.05,1471-2} = 1.65$. Since $|t| > t_{0.05,1471-2}$, we reject H_0.
- $p < 10^{-16}$, and 95% confidence interval is (157.18, 242.29).

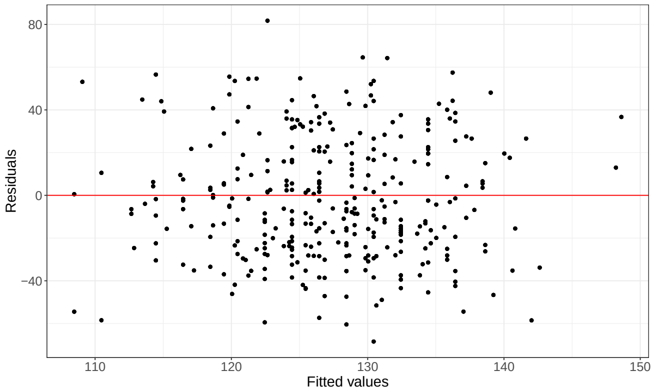
Assumptions in linear regression

- Assume an underlying linear relationship between the two variables x and y.
- Normality: Assume for a given value of x, that y follows a normal distribution.
- Independence of observations.
- Homoscedasticity: variance of y does not change over the values of x.

Residual plot

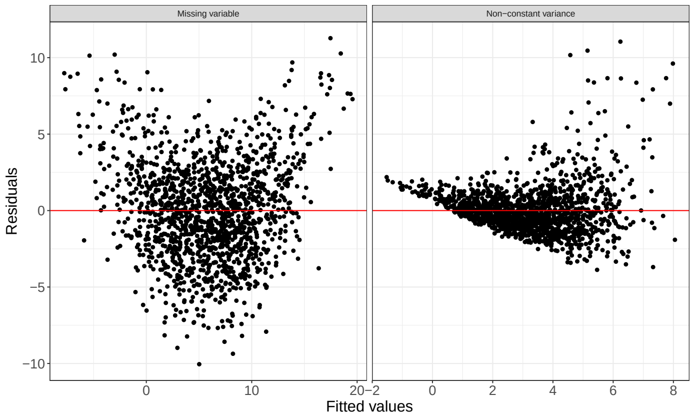
- Looking at residuals is an excellent way to check model assumptions
- Most basic tool: plotting the fitted values vs. the residuals
- Don’t want to see any kind of discernible pattern in the residual plot. Otherwise:
    - Could have non-constant variance
    - Could have important variables missing
- Can also check the distribution of the residuals to see if normal distribution assumption is met.

IMPC data residual plot



Check variance of residuals over the fitted values. In this example, there is no discernible pattern.

Bad residual plots

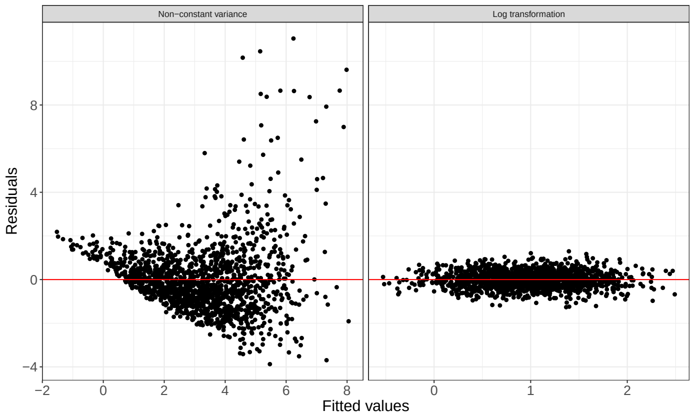


Two examples of bad residual plots (simulated data).

Action for bad residual plot

- Sometimes have to experiment a little.
- Can add extra variables to model (multiple linear regression).
- Could add additional higher order terms to model.
  - E.g. If age is in the predictor, could add age^2 as an additional predictor in the model.
- Could do transformations on response variable to get constant variance: log, square root.
  - Careful... this changes the interpretation of b_1.
- Don't look at the p-value during this process!

Fixed residual plot



Original residual plot on left. Fixed residual plot corresponding to log-transformed response variable on right.

Multiple Regression

Multiple regression formulation

- Multiple regression is very similar to simple linear regression. But now there is more than one predictor variable (still a single response variable).
- E.g. if there were three predictor variables $x_1 = (x_{11}, x_{12}, \ldots, x_{1n}), x_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, $x_3 = (x_{31}, x_{32}, \ldots, x_{3n})$, then the regression model would be:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \epsilon_i$$

- Slope parameters b_1, b_2, and b_3 measure association between the predictors and y.
- All assumptions from before still present (linearity, normality, independence, constant variance)

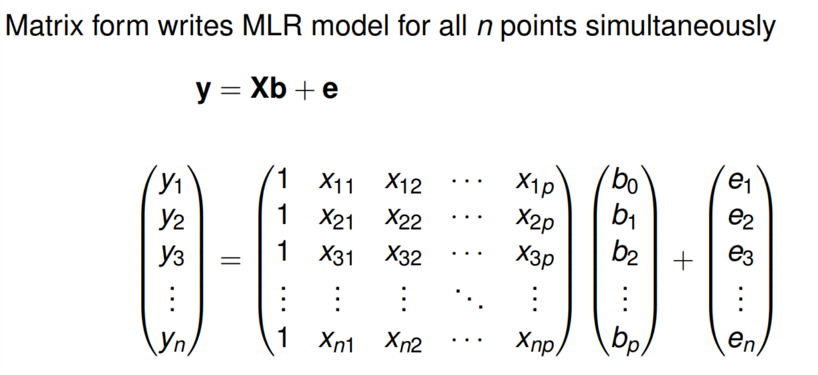
Why use multiple regression?

- Could be interested in the joint effect of multiple variables on a single outcome variable.
    - Estimated effects are different than effects for a separate linear regression model run for each predictor.
- Even if only interested in one predictor and outcome, can included additional variables in model to “adjust” and therefore reduce potential bias.
    - E.g. in our total cholesterol vs. creatinine example, we could include mouse body weight as an additional variable in the model.
    - The estimated association between cholesterol and creatinine would then be adjusted for body weight.
- Including more variables in the model can often improve efficiency in estimates of association (smaller standard errors).

Interpretation of slopes

- Slope parameters b_1, b_2, and b_3 are of interest, but the interpretations are a bit different.
- $\hat{b}_1$ is the estimated increase in the response y associated with an increase in one unit of x_1 when **all other variables are held constant**.
- Likewise interpretations for b_2 and b_3

Data Setup



Matrix form writes MLR model for all n points simultaneously

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

Multiple Linear Regression Objective Function

If we have p predictors, then we need to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ parameters where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Using the least-squares criterion

$$\widehat{\boldsymbol{\beta}} = \arg \min_{(\beta_0, \beta_1, \ldots, \beta_p)} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)$$

the solution to this objective function is given by

$$\widehat{\boldsymbol{\beta}} = (\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{X}^\top \mathsf{Y}$$

with variance given by

$$Var(\widehat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathsf{X}^\top \mathsf{X})^{-1}$$

where $\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Hypothesis testing

- Can do individual hypothesis tests for regression parameters. For each parameter $j \in \{1, 2, 3\}$:

  $$H_0 : b_j = 0$$
  $$H_1 : b_j \neq 0$$

- Testing the individual parameters once again results in t-tests. The test statstic is:

  $$t = \frac{\hat{b}_j}{SE(\hat{b}_j)}$$

IMPC example

- Consider looking at the how total cholesterol changes with respect to creatinine, glucose, and body weight.
- Regression model:

$$y_i = b_0 + b_1\text{creatinine} + b_2\text{glucose} + b_3\text{weight}$$

	Estimate	Std. Error	*t*-value	*p*-value
creatinine	67.91	19.71	3.45	5.86×10^{-4}
glucose	0.14	0.01	12.49	4.10×10^{-34}
weight	3.44	0.14	24.74	3.32×10^{-113}

Fitted model

- Can rewrite the fitted model as:

$$\hat{y}_i = -25.40 + 67.91 \times \text{creatinine}_i + 0.14 \times \text{glucose}_i$$
$$+ 3.44 \times \text{weight}_i$$

Take-home message

- Regression is a very powerful and versatile tool.
- Do a thorough investigation of model assumptions.
- Many assumptions to make, but lots of other models exist if assumptions are not met
    - Non-constant variance: Weighted least squares
    - Non-normal data: generalized linear models
    - Observations not independent: random effects models
- A lot of room for choosing models. Make your choice based on good statistical principles... not on the resulting *p*-values!

Session Info

```r
devtools::session_info()
```

```
##  setting  value
##  version  R version 3.4.1 (2017-06-30)
##  system   x86_64, linux-gnu
##  ui       X11
##  language en_US
##  collate  en_US.UTF-8
##  tz       Canada/Eastern
##  date     2017-11-28
##
##  package     * version    date       source
##  abind         1.4-5      2016-07-21 cran (@1.4-5)
##  arm           1.9-3      2016-11-27 cran (@1.9-3)
##  assertthat    0.2.0      2017-04-11 CRAN (R 3.4.1)
##  backports     1.1.0      2017-05-22 cran (@1.1.0)
##  base        * 3.4.1      2017-07-08 local
##  bindr         0.1        2016-11-13 CRAN (R 3.4.1)
##  bindrcpp      0.2        2017-06-17 CRAN (R 3.4.1)
##  blme          1.0-4      2015-06-14 cran (@1.0-4)
##  broom         0.4.2      2017-02-13 CRAN (R 3.4.1)
```