Linear regression using ordinary least squares

Malachy Campbell

Sept. 4, 2020

Linear models

Is there a relationship between body weight and height?

- What is the magnitude of this association? (estimation)
- How tall is someone given their body weight? (prediction)

Some review...

Variance Expectation:

$$\sigma_x^2 = E[(x - \mu_x)^2]$$

Sampling estimator:

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Covariance Expectation:

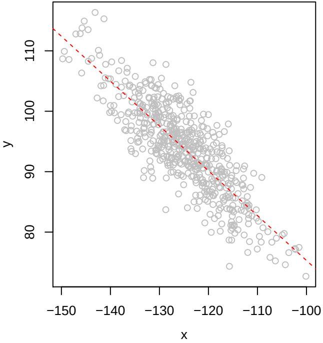
$$\sigma_{x,y}^2 = E[(x - \mu_x)(y - \mu_y)]$$

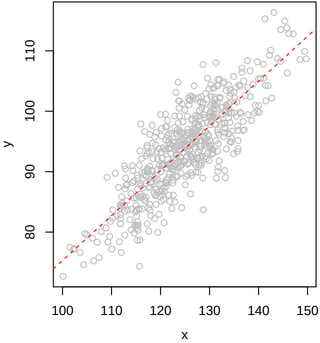
Sampling estimator::

$$\hat{\sigma}_{x,y} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Some review...

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$





$r = 0.89$ $r = -0.89$

Some more review...

- Scalar: single real number
- Vector: collection of scalars
  - A slice of a matrix
- Matrix collection of vectors

A vector:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

A matrix:

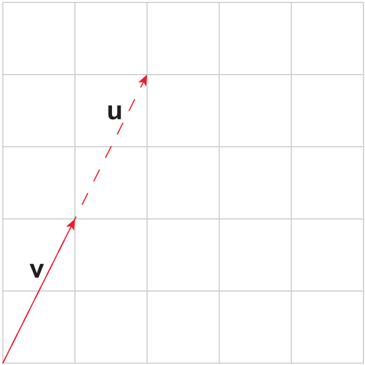
$$\begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}$$

Some more review...

Multiplying vector by scalar

$$\Theta = 2, \mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$\mathbf{u} = \Theta \mathbf{v} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

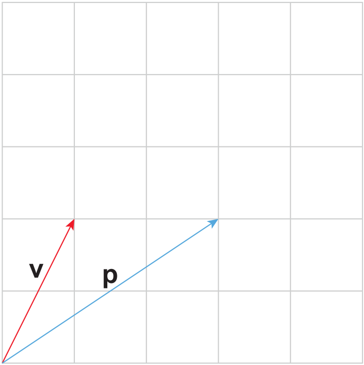


Some more review...

Vector addition (must have same no. elements)

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{p} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$$

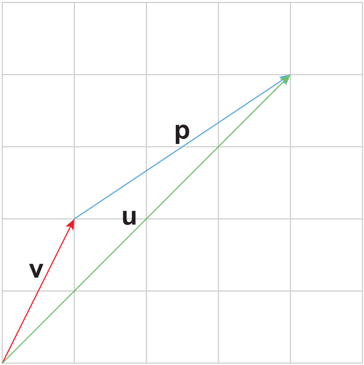


Some more review...

Vector addition (must have same no. elements)

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{p} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$$



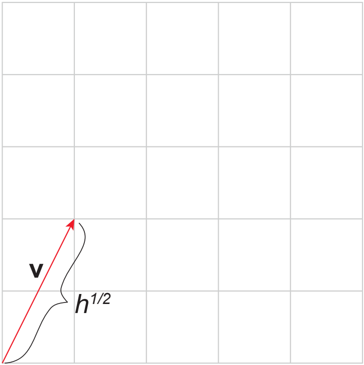
Some more review...

Inner product (returns scalar)

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

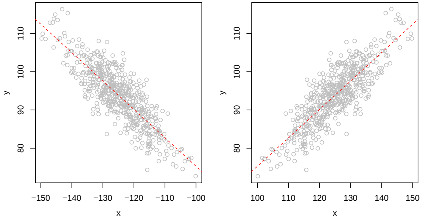
$$h = \mathbf{v}'\mathbf{v} = \sum_i v_i v_i = 5$$

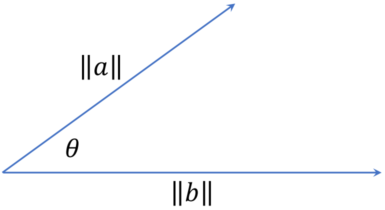
$$h = \mathbf{v}'\mathbf{v} = \sum_i v_i v_i = ||\mathbf{v}||^2$$



A geometric interpretation of correlation

What is the strength and direction of the relationship between the vectors x and y?





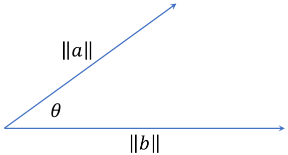
$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

$$\mathbf{a} = \bar{\mathbf{x}} - \mathbf{x}$$
$$\mathbf{b} = \bar{\mathbf{y}} - \mathbf{y}$$

$$r_{x,y} = cos(\theta)$$

A geometric interpretation of correlation

The cosine formula for the dot (inner) product:



$$r_{x,y} = cos(\theta)$$

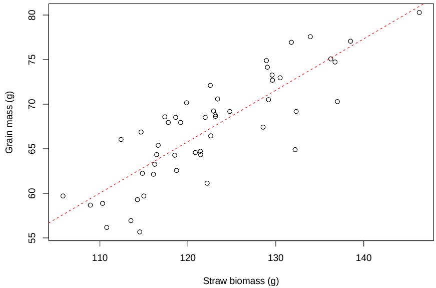
$$
\begin{aligned}
\mathbf{a}'\mathbf{b} &= \|\mathbf{a}\|\|\mathbf{b}\|cos\theta \\
\frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} &= cos\theta \\
\frac{\mathbf{a}'\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{b}}} &= \\
\frac{\mathbf{a}'\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{b}}} &= \\
\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} &= \\
\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} &=
\end{aligned}
$$

Least squares regression – Example

We are interested in looking at the relationship between straw biomass and bield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal is to find a line defined by β_0 and β_1 that best fit the data

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



Least squares regression – Example

We are interested in looking at the relationship between straw biomass and bield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal is to find a line defined by β_0 and β_1 that best fit the data

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

- Response y_i: Yield for the ith experimental unit
- β_0: intercept (overall mean)
- β_1: deviation from overall mean due to biomass
- x_i: straw biomass for the ith experimental unit
- e_i random deviation from known experimental factors

Least squares regression – Example

We are interested in looking at the relationship between straw biomass and bield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal is to find a line defined by β_0 and β_1 that best fit the data

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

- The "best fit" is one where the difference between the predicted ($\hat{\mathbf{y}} = \beta_0 + \beta_1 \mathbf{x}$) and observed values are smallest
- The "best fit" is one that minimizes the sum of the squared residuals
    - Residuals: $\hat{\mathbf{e}} = \hat{\mathbf{y}} - \mathbf{y}$

Least squares regression – Example

We are interested in looking at the relationship between straw biomass and bield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal is to find a line defined by β_0 and β_1 that best fit the data

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

Alternatively, these can be represented as vectors.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Least squares regression – Example

We are interested in looking at the relationship between straw biomass and bield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal is to find a line defined by β_0 and β_1 that best fit the data

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

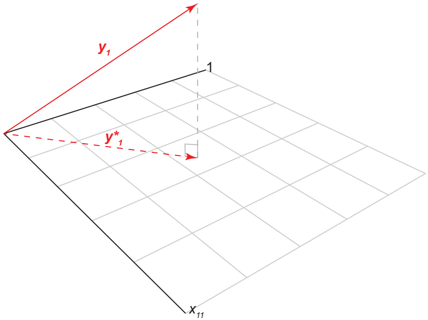
Grouping the predictor vectors into a matrix

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Geometric interpretation of least squares

- Our data can be represented as vectors that exist in a 3d space
- All predictors exist in a 2d plane

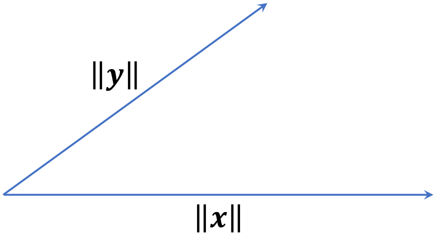
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}$$



Geometric interpretation of least squares

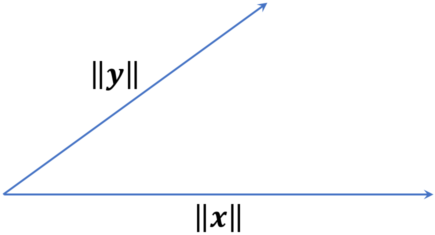
- Our data can be represented as vectors that exist in a 3d space
- All predictors exist in a 2d plane
- If we center the response and predictors (subtract mean of vector from their elements) the intercept drops out

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}$$



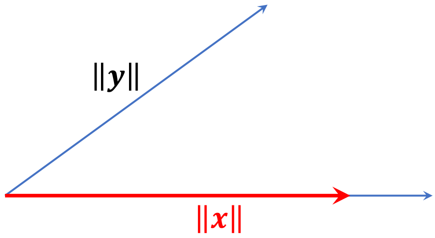
Geometric interpretation of least squares

- Find the shortest vector from $\|y\|$ that intersects the plane of $\|x\|$



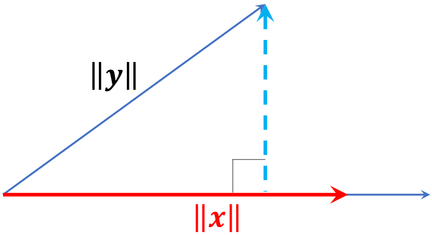
Geometric interpretation of least squares

- Find the shortest vector from $\|y\|$ that intersects the plane of $\|x\|$



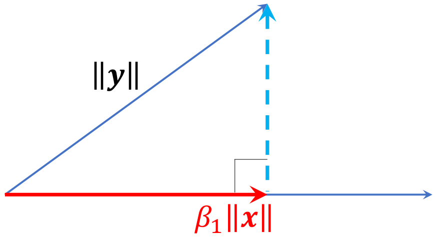
Geometric interpretation of least squares

- Find the shortest vector from $\|y\|$ that intersects the plane of $\|x\|$
- We need to find some value, $\hat{\beta}_1$, that when multiplied by $\|x\|$ gives a vector that joins all three vectors.



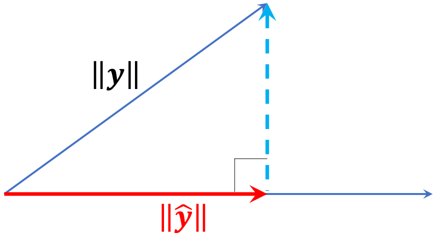
Geometric interpretation of least squares

- Find the shortest vector from $\|y\|$ that intersects the plane of $\|x\|$
- We need to find some value, $\hat{\beta}_1$, that when multiplied by $\|x\|$ gives a vector that joins all three vectors.



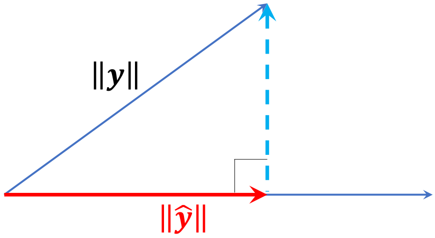
Geometric interpretation of least squares

- Find the shortest vector from $\|y\|$ that intersects the plane of $\|x\|$
- We need to find some value, $\hat{\beta}_1$, that when multiplied by $\|x\|$ gives a vector that joins all three vectors.
- How can we find the length of the blue vector (c) given $\|\hat{\mathbf{y}}\|$ and $\|\mathbf{y}\|$?



Geometric interpretation of least squares

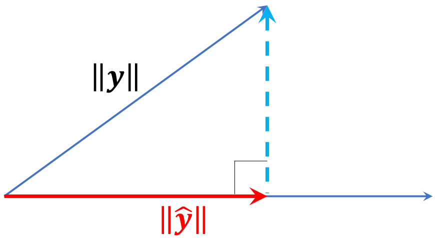
- Find the shortest vector from $\|y\|$ that intersects the plane of $\|x\|$
- We need to find some value, $\hat{\beta}_1$, that when multiplied by $\|x\|$ gives a vector that joins all three vectors.
- How can we find the length of the blue vector (c) given $\|\hat{\mathbf{y}}\|$ and $\|\mathbf{y}\|$?



$$
\begin{aligned}
\|\mathbf{y}\|^2 &= c + \|\hat{\mathbf{y}}\|^2 \\
\|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2 &= c \\
\|\mathbf{y} - \hat{\mathbf{y}}\|^2 &= c
\end{aligned}
$$

Geometric interpretation of least squares

$$\|\mathbf{y}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}}\|^2$$
$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{e}}\|^2 + \|\hat{\mathbf{y}}\|^2$$



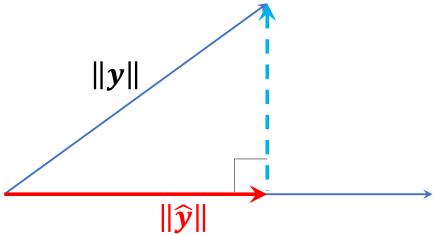
Geometric interpretation of least squares

$$\begin{aligned}\|\mathbf{y}\|^2 &= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}}\|^2 \\ \|\mathbf{y}\|^2 &= \|\hat{\mathbf{e}}\|^2 + \|\hat{\mathbf{y}}\|^2\end{aligned}$$

Since

$$\|\mathbf{v}\|^2 = \sum_{i=1}^{n} v_i^2$$

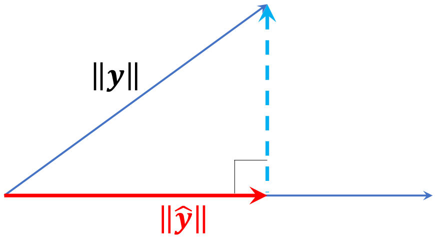
$$\begin{aligned}\|\mathbf{y}\|^2 &= \|\hat{\mathbf{e}}\|^2 + \|\hat{\mathbf{y}}\|^2 \\ \sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} \hat{e}_i^2 + \sum_{i=1}^{n} \hat{y}_i^2\end{aligned}$$



Geometric interpretation of least squares

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{e}_i^2 + \sum_{i=1}^{n} \hat{y}_i^2$$

- With least squares we are partitioning the total sum of squares ($\sum_{i=1}^{n} y_i^2$) into sum of squares from our regression model ($\sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} (\beta_1 x_i)^2$) and the error sum of squares ($\sum_{i=1}^{n} e_i^2$).
  - **The regression sum of squares and error sum of squares are orthogonal ($cos(90) = 0$)**



Least squares

We are interested in looking at the relationship between straw biomass and yield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal: Find a line $\beta_0 + \beta_1 x$ that best approximates this relationship The

"best" fit is one that minimizes the sum of the squared residuals.

Least squares

We are interested in looking at the relationship between straw biomass and yield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.

Goal: Find a line $\beta_0 + \beta_1 x$ that best approximates this relationship The

"best" fit is one that minimizes the sum of the squared residuals. Ordinary least squares cost function:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} e_i^2$$

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} (y_i - \beta_0 + \beta_1 x_i)^2$$

Derivation of OLS estimators

Cost function:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} (y_i - \beta_0 + \beta_1 x_i)^2$$

Set partial derivatives to 0:

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

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

Derivation of OLS estimators

Solve for intercept:

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

$$N\bar{y} - N\beta_0 - N\beta_1\bar{x} = 0$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

Derivation of OLS estimators

Solve for slope: We know

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

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

$$\sum_{i=1}^{N} x_i y_i - (\bar{y} - \beta_1 \bar{x})x_i - \beta_1 x_i^2 = 0$$

$$\sum_{i=1}^{N} x_i y_i - N\bar{y}\bar{x} + N\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^{N} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}$$

Derivation of OLS estimators

Solve for slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}$$

OR

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Thus, the slope is the the covariance of x and y relative to the variance of x.

Some more matrix algebra review...

To transpose a matrix we simply swap the rows and columns.

- Transpose is denoted using ^{T} or $'$.

$$\mathbf{C} = \begin{bmatrix} 1 & 2 \\ 4 & 7 \end{bmatrix}$$

$$\mathbf{C}^T = \begin{bmatrix} 1 & 4 \\ 2 & 7 \end{bmatrix}$$

Some more matrix algebra review...

$$\mathbf{C} = \mathbf{AB}$$

- Split each matrix into vectors
  - First matrix is split into row vectors
  - Second matrix is split into column vectors
- Each element of the resulting matrix is then calculated via:

$$C_{i,j} = \mathbf{a}_i \cdot \mathbf{b}_j = \sum_{k=1}^{c} a_{ik} \cdot b_{jk}$$

- i row, j column
- c is the number of columns in $\mathbf{A}$ = number of rows in $\mathbf{B}$

Regression in matrix form

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

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

Simple linear regression:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Regression in matrix form

Assume we centered the response and predictor variables. The OLS solution in matrix form becomes

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- $\mathbf{X}'\mathbf{X}$: (Co)variance of x
- $\mathbf{X}'\mathbf{y}$: covariance of x and y

Regression Diagnostics

- Is the model valid? Are we violating any assumptions?
- Are any points overly influential?
- Are there any outliers?

Regression Diagnostics

- Is the model valid? Are we violating any assumptions?
- Are any points overly influential?
- Are there any outliers?

Assumptions:

- **Linearity**: $\mathrm{E}[y|x]$ is linearly related to x
- **Independence**: observations are indep. of one another
- **Normality**: Dist. of $[y|x]$ is normal
- **Equal variance**: $\mathrm{Var}[y|x]$ is not dependant on the value of x

Regression Diagnostics – hat (or projection) matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Hat matrix (H): Links predicted values with response variable.

- Projection because it projects the vector $\mathbf{y}$ into the space of $\mathbf{X}$
  - $\hat{\mathbf{y}} = \boldsymbol{X}\beta = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$
- Gives insight into the "influence" of an observation
- Is used to compute standardized residuals

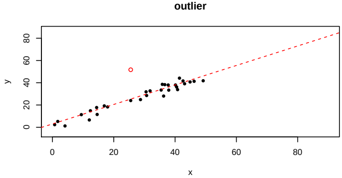
Outliers vs leverage points

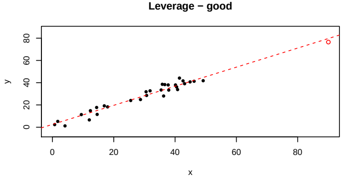
Outlier:

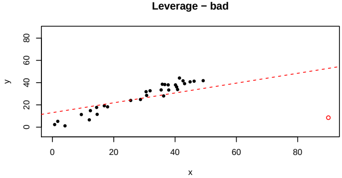
- A point that deviates from the pattern set by rest of data

Leverage point:

- A point with an x-value that is far from rest of x-values and "pulls" regression line towards point
- *Good* if the y-value follows true regression line or OLS line
- *Bad* if the y-value deviates from true regression line or and affects OLS line







Leverage Points

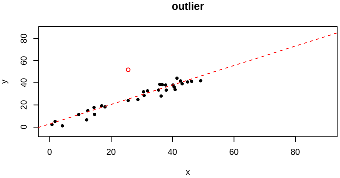
Diagonal elements of the projection matrix...

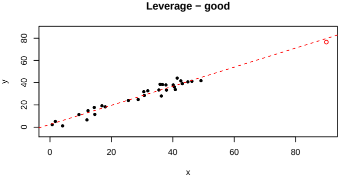
$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

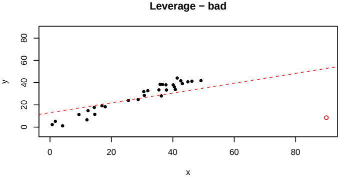
...or for a given point

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

$$\sum_i h_i = 1$$



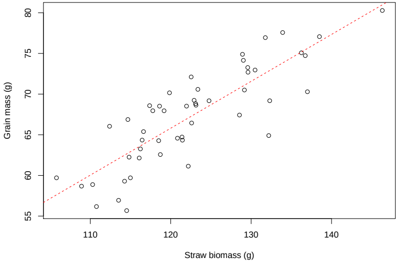




Is the model valid?

- **Linearity**: $\mathrm{E}[y|x]$ is linearly related to x
- **Independence**: observations are indep. of one another
- **Normality**: Dist. of $[y|x]$ is normal
- **Equal variance**: $\mathrm{Var}[y|x]$ is not dependant on the value of x

Linear relationship between x and y:



Regression diagnostics

Most diagnostics are based on looking at the behavior of the residuals

Recall,

$$e_i = y_i - \hat{y}_i$$

And the unbiased estimate of the population variance is given by

$$\hat{\sigma}^2 = MSE = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

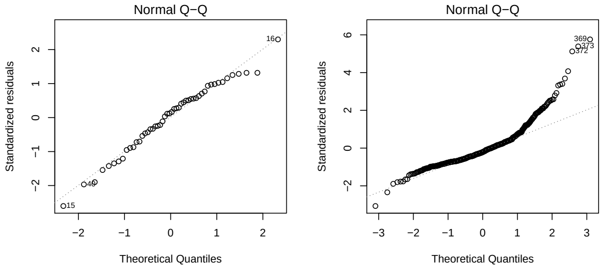
The residuals can be standardized by dividing by the standard deviation

$$z_i = \frac{e_i}{\sqrt{MSE}}$$

Regression diagnostics – Normality

$$z_i = \frac{e_i}{\sqrt{MSE}}$$

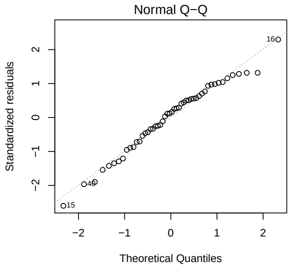
Under the assumptions of the linear model $z_i \sim N(0, 1)$. Thus we can compare the distribution of z_i with the quantiles of a standard normal distribution.

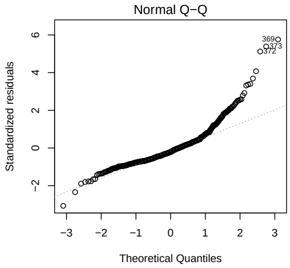


Regression diagnostics – Normality

$$z_i = \frac{e_i}{\sqrt{MSE}}$$

Under the assumptions of the linear model $z_i \sim N(0, 1)$. Thus we can compare the distribution of z_i with the quantiles of a standard normal distribution.

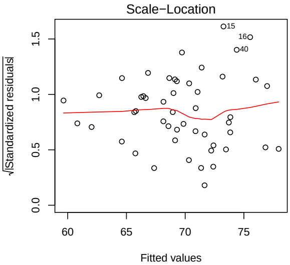


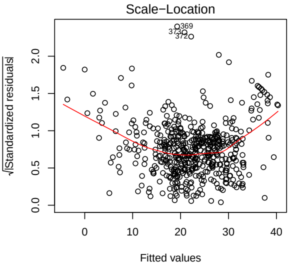


Not particularly a problem with large datasets

Regression Diagnostics – equal (constant) variance

$\sqrt{|Std.Residuals|}$ vs fitted values

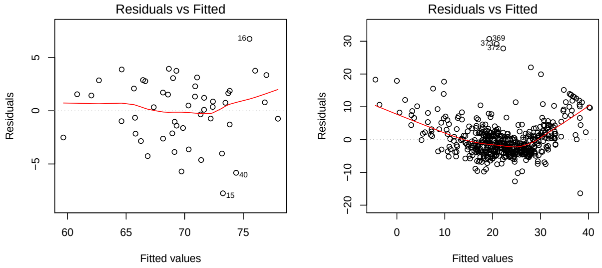




- Red line should be somewhat flat and horizontal (linearity) and show no obvious "spread" with fitted values (homoscedasticity) (left)
  - Square root is used to reduce skewness of std. residuals

Regression Diagnostics – equal (constant) variance

Residuals vs fitted values



- Points should follow a horizontal line and the mean of the residuals should be close to 0 (left)
  - If the |residuals| gets larger as fitted values get larger then variance is likely not constant
  - If there are convex or concave patterns then x and y are likely not linear

Lack of fit

How well does our model fit the data?

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$
$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

- SST: Total sum of squares
- SSR: Regression sum of squares
  - How far fitted values are from the mean
- SSE: Error/residual sum of squares
  - How far the fitted values are from the observations

Lack of fit

How well does our model fit the data?

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$
$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

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

Testing for lack of fit

Source	**SS**	**df**	**MS**
Regression	$SSR = \sum_{i=1}^{i} (\hat{y}_i - \bar{y})^2$	$df_{Reg.} = p$	$\frac{SSR}{df_{Reg.}}$
Error	$SSE = \sum_{i=1}^{i} (y_i - \hat{y}_i)^2$	$df_{Err.} = n - p - 1$	$\frac{SS_{Err.}}{df_{Err.}}$
Total	$SS_{Tot.} = \sum_{i=1}^{i} \sum_{j=1}^{j} (y_i - \bar{y})^2$	$df_{Tot.} = n - 1$	

- p is the number of predictors in the model, not including the intercept

Testing for lack of fit

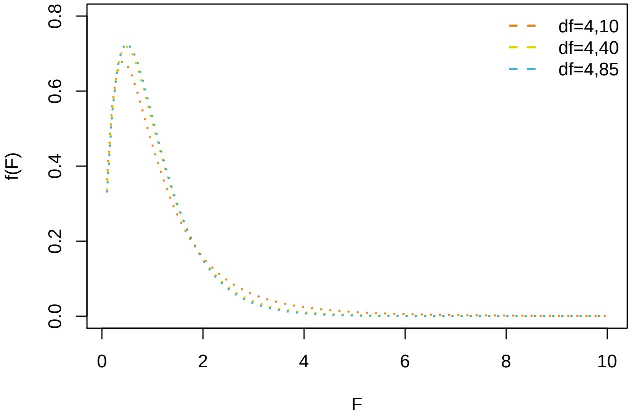
Source	**SS**	**df**	**MS**
Regression	$SSR = \sum_{i=1}^{i}(\hat{y}_i - \bar{y})^2$	$df_{Reg.} = p$	$\frac{SSR}{df_{Reg.}}$
Error	$SSE = \sum_{i=1}^{i}(y_i - \hat{y}_i)^2$	$df_{Err.} = n - p - 1$	$\frac{SS_{Err.}}{df_{Err.}}$
Total	$SS_{Tot.} = \sum_{i=1}^{i}\sum_{j=1}^{j}(y_i - \bar{y})^2$	$df_{Tot.} = n - 1$	

- p is the number of predictors in the model, not including the intercept
- We can test if there is a good fit, by using an F-test
  - The F-test (or F-statistic) is used to determine whether the variance explained by the model is significantly more than the error variance (more on this next lecture...)

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

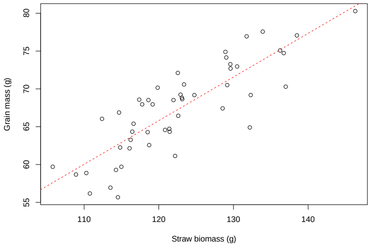
Testing for lack of fit

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$



Regression example

We are interested in looking at the relationship between straw biomass and yield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.



Predictors:

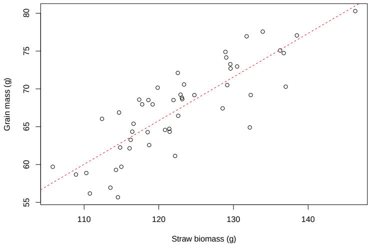
```
> Xincd <- model.matrix(~ 1 + dataSet$BM) # predictors (intercept and biomass)
> head(Xincd)
  (Intercept) dataSet$BM
1           1   136.4120
2           1   124.1772
3           1   129.2626
4           1   120.3829
5           1   124.7200
6           1   125.1760
```

OLS estimates:

```
> Y <- dataSet$Yld
> solve(t(Xincd) %*% Xincd) %*% t(Xincd) %*% Y
                 [,1]
(Intercept) 2.1967411
dataSet$BM  0.5300663
```

Regression example

We are interested in looking at the relationship between straw biomass and yield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.



How well does our model fit the data?

```
> anova(lm(Yld ~ BM, dataSet))
Analysis of Variance Table

Response: Yld
          Df Sum Sq Mean Sq F value    Pr(>F)
BM         1 904.00  904.00  97.396 3.894e-13 ***
Residuals 48 445.52    9.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Inference on β's

We can also perform tests to see if any of the coefficients are significant different from some specified value (most often 0).

- **Hypothesis:** H_0: $\beta_1 = 0$; H_0: $\beta_1 \neq 0$;
- **Test stat.:** signal to noise ratio

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

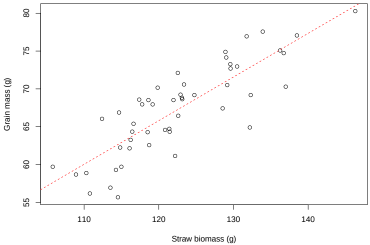
- Null sampling distribution: t-distribution with $n - p$ d.f.

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \frac{\mathbf{e}'\mathbf{e}}{n-p}(\mathbf{X}'\mathbf{X})^{-1}$$

- The square root of the diagonal gives the standard errors of the coefficient estimates.

Regression example – Inference on β's

We are interested in looking at the relationship between straw biomass and yield in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass and yield were collected.



Equivalent to F-test since we have one predictor.

```
> summary(lm(Yld ~ BM, dataSet))

Call:
lm(formula = Yld ~ BM, data = dataSet)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0551 -1.5520  0.4199  2.2288  6.2904 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.19674    6.75943   0.325    0.747    
BM           0.53007    0.05371   9.869 3.89e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.047 on 48 degrees of freedom
Multiple R-squared:  0.6699,	Adjusted R-squared:  0.663 
F-statistic:  97.4 on 1 and 48 DF,  p-value: 3.894e-13
```

Prediction

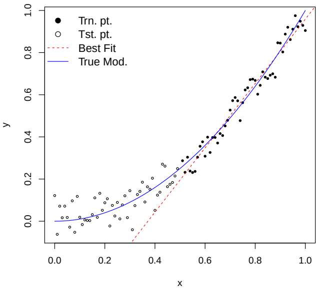
- Point estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Standard error of estimate (expectation for all indiv. with $x = x_i$):

$$\mathrm{SE}(\hat{y}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} y_i - \hat{y}}{n-2}}$$

Beware: Predicting outside of range of data (i.e. extrapolation)

- Note that the SE of prediction is dependant on how far predictor is from mean of predictors.



Prediction

- Point estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Standard error of estimate (expectation for all indiv. with $x = x_i$):

$$\mathrm{SE}(\hat{y}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

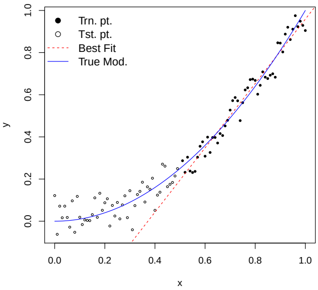
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} y_i - \hat{y}}{n-2}}$$

- Standard error of estimate (single indiv. with $x = x_i$):

$$\mathrm{SE}(\hat{y^*}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Beware: Predicting outside of range of data (i.e. extrapolation)

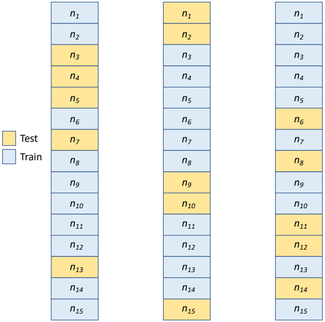
- Note that the SE of prediction is dependant on how far predictor is from mean of predictors.



How accurate are my predictions?

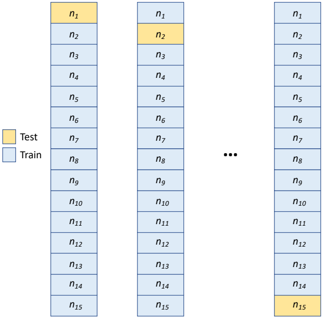
Cross validation: Split dataset, train on subset, and predict remaining observations

K-fold cross validation:



$$r = cor(\hat{y}_{tst}, y_{tst})$$

Leave-one-out cross validation:



$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$