
Lecture 3: Simple Linear Regression

Reading

- Weisberg, Chapter 2: “Simple Linear Regression” (Required)

We will consider the case in which we observe a single response and one covariate. (For most of the course, we assume that the covariates are fixed and known.)

This lecture will serve as a gentle introduction to the matrix-based material utilized in the rest of this course, and provide some intuition and motivations for later concepts.

Model Notation

- Roman letters $(a, b, c, \dots, x, y, z)$ represent constants and random variables.
 - Letters at beginning of the alphabet, (a, b, c, \dots) , often represent constants.
 - Letters in the middle of the alphabet, (\dots, i, j, k, \dots) , often represent indices.
 - Letters at the end of the alphabet, (\dots, x, y, z) , often represent random variables.
- Greek letters $(\alpha, \beta, \gamma, \dots, \chi, \psi, \omega)$ represent parameters to be estimated.
- Lower-case symbols in bold type $(\mathbf{x}, \mathbf{y}, \boldsymbol{\beta}, \dots)$ represent vectors.
- Upper-case symbols in bold type $(\mathbf{X}, \mathbf{A}, \boldsymbol{\Sigma}, \dots)$ represent matrices.

Simple Linear Regression

Regression analysis is used to explore the nature of the relationship between a response variable and one or more covariates. In simple linear regression, we focus on the case when we consider only a single covariate.

Hypothesis: State-specific melanoma mortality rates are related to the latitude of the state.

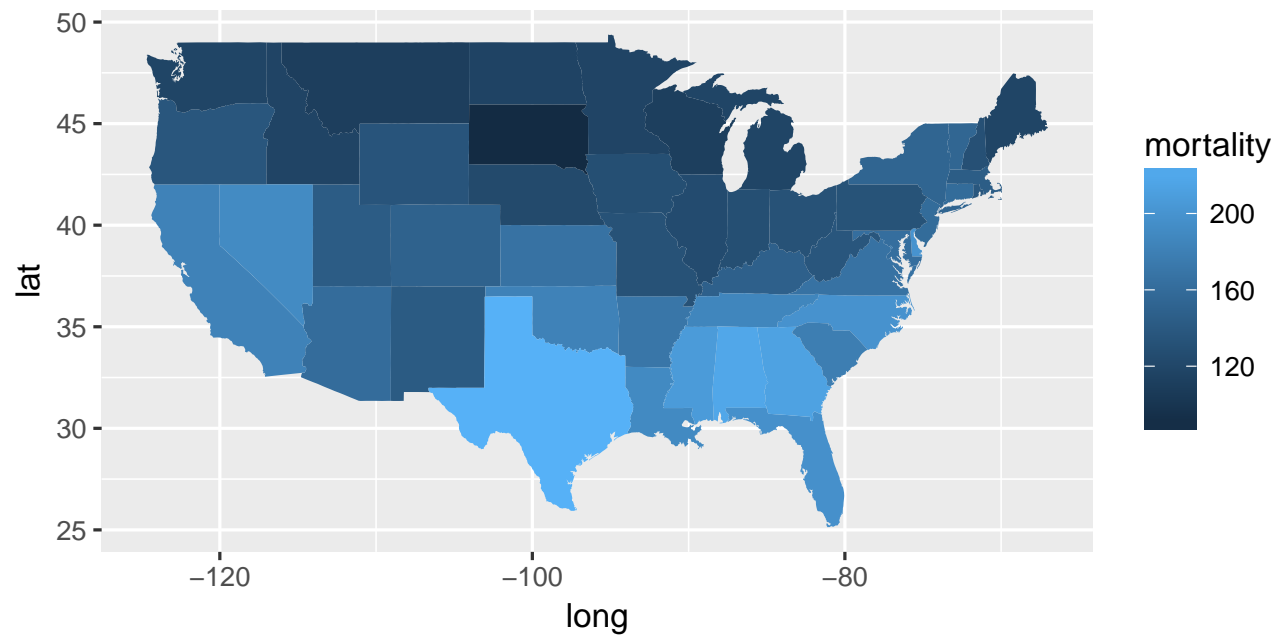
Define:

y_i = annual melanoma mortality in state i , $i = 1, \dots, 50$

x_i = latitude of the center of the state (in degrees) 

Our observations are $(x_1, y_1), \dots, (x_{50}, y_{50})$, where the x_i 's are known fixed values (*predictors, covariates, or independent variables*), and the y_i 's are random *response variables (dependent variables)*.

<i>State</i>	<i>Melanoma Mortality Rate</i> (Deaths/Million)	<i>Latitude</i> (in degrees)
Alabama	219	33.00
Alaska	220	63.25
Florida	197	28.00
Hawaii	330	19.96
Maine	117	45.20
Minnesota	116	46.00
Mississippi	207	32.80
North Dakota	115	47.50
North Carolina	199	35.50
Tennessee	186	36.00
Vermont	153	44.00




How we do determine the relationship between melanoma mortality
and latitude?

Simple Linear Regression

Let us write the simple linear regression (SLR) model as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n \quad (1)$$

where

- n is the number of observations or sampling units
- y_1, \dots, y_n are the random responses,
- x_1, \dots, x_n are the fixed and known covariates
- $\varepsilon_1, \dots, \varepsilon_n$ is a vector of unobserved random errors, s.t.
 $\varepsilon_i \stackrel{i.i.d}{\sim} N(0, \sigma^2)$ 
- β_0, β_1 , and σ^2 are parameters (fixed and to be estimated from the data)




The index i corresponds to subjects or sampling units.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$$

where

- $E[\varepsilon_i | x_i] = E[\varepsilon_i] = 0$
- On right side, only ε_i is random
- (y_i, x_i) forms the data for observation i

Model Assumptions

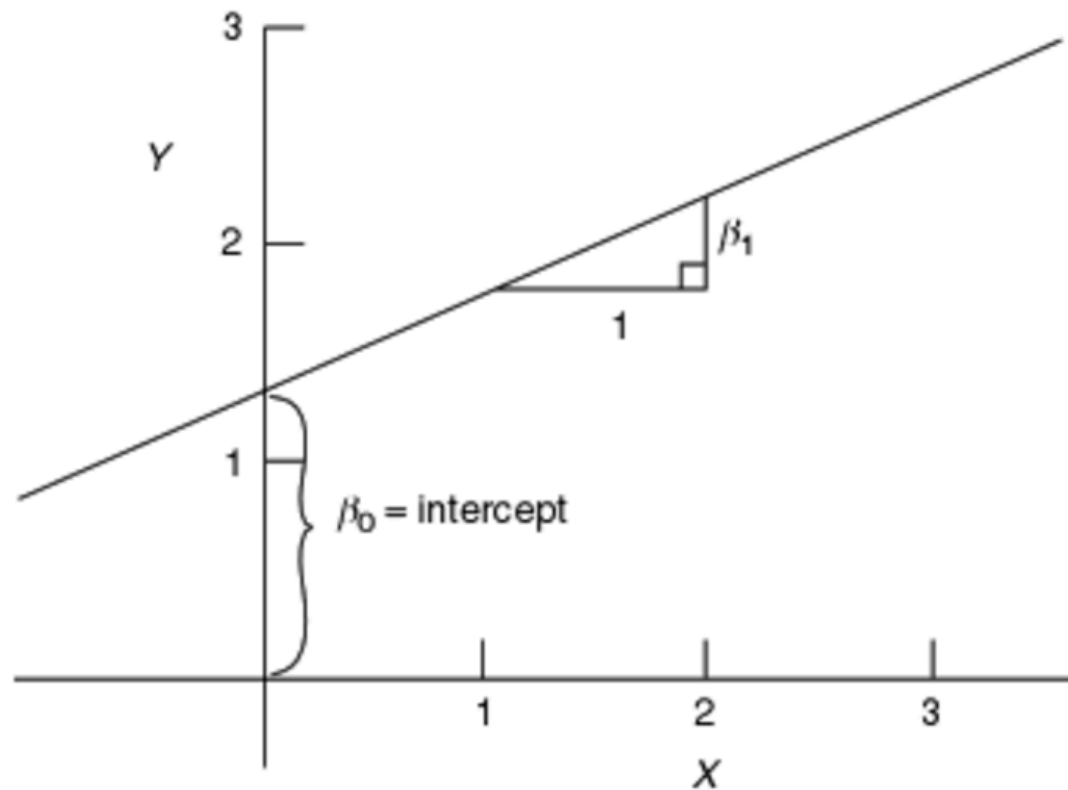
- Homogeneity 
- Independence 
- Linearity
- Existence 
- Gaussian Errors

Violations of these assumptions lead to problems with the application, estimation, and interpretation of linear models.

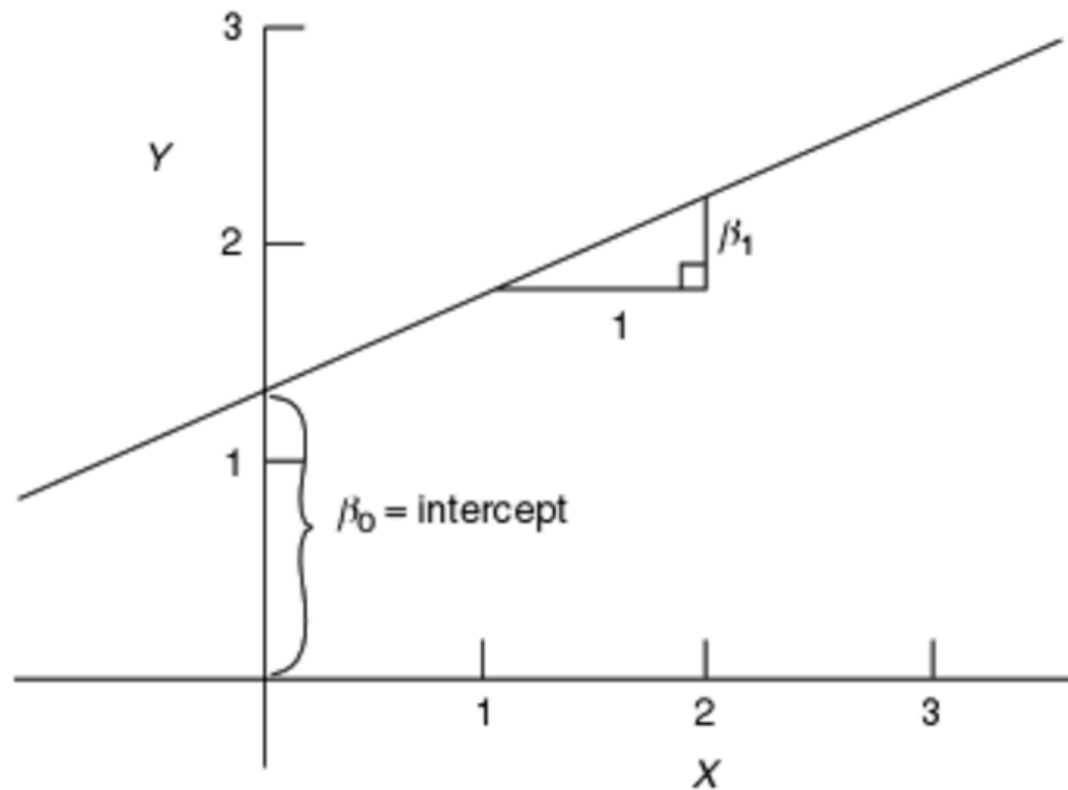
We also may write the simple linear regression (SLR) model by taking expectations of both sides

$$\begin{aligned}E[y_i|x_i] &= E[\beta_0 + \beta_1 x_i + \epsilon_i|x_i] \\&= \beta_0 + \beta_1 x_i \\Var[y_i|x_i] &= Var[\beta_0 + \beta_1 x_i + \epsilon_i|x_i] \\&= \sigma^2\end{aligned}$$

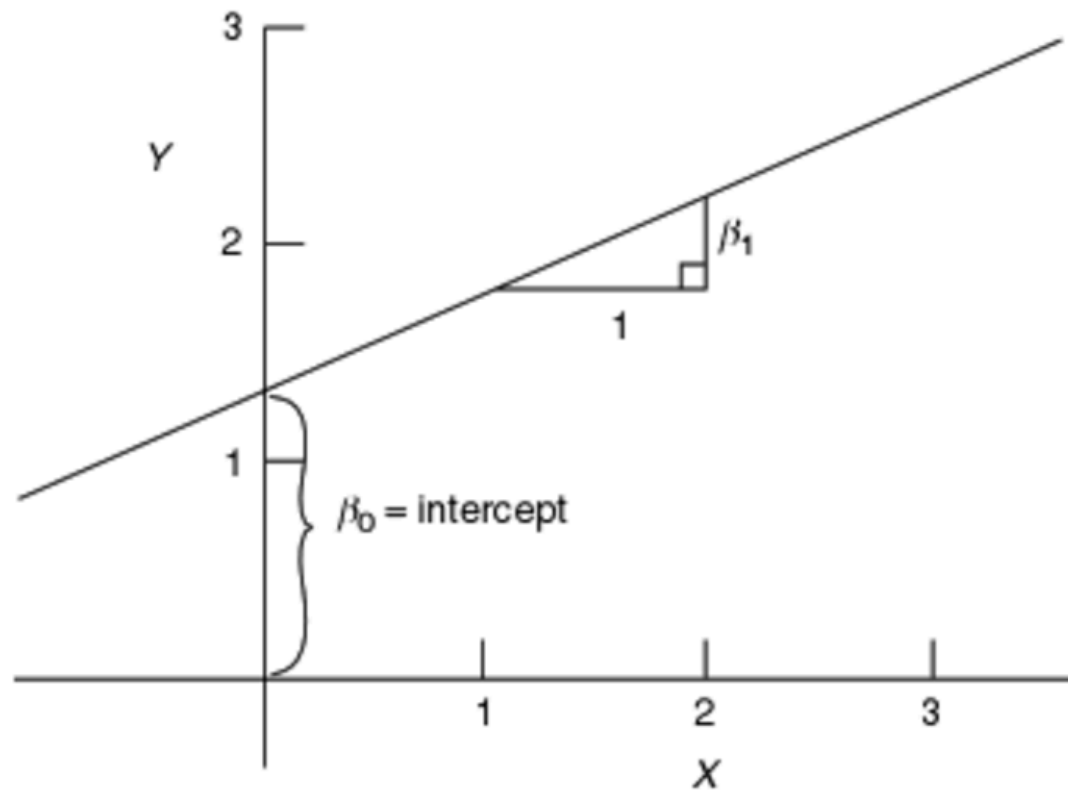
Based on the assumptions of the SLR, how did we arrive at this?
What does this say about the distribution of y_i ? We can plot this line on the next slide



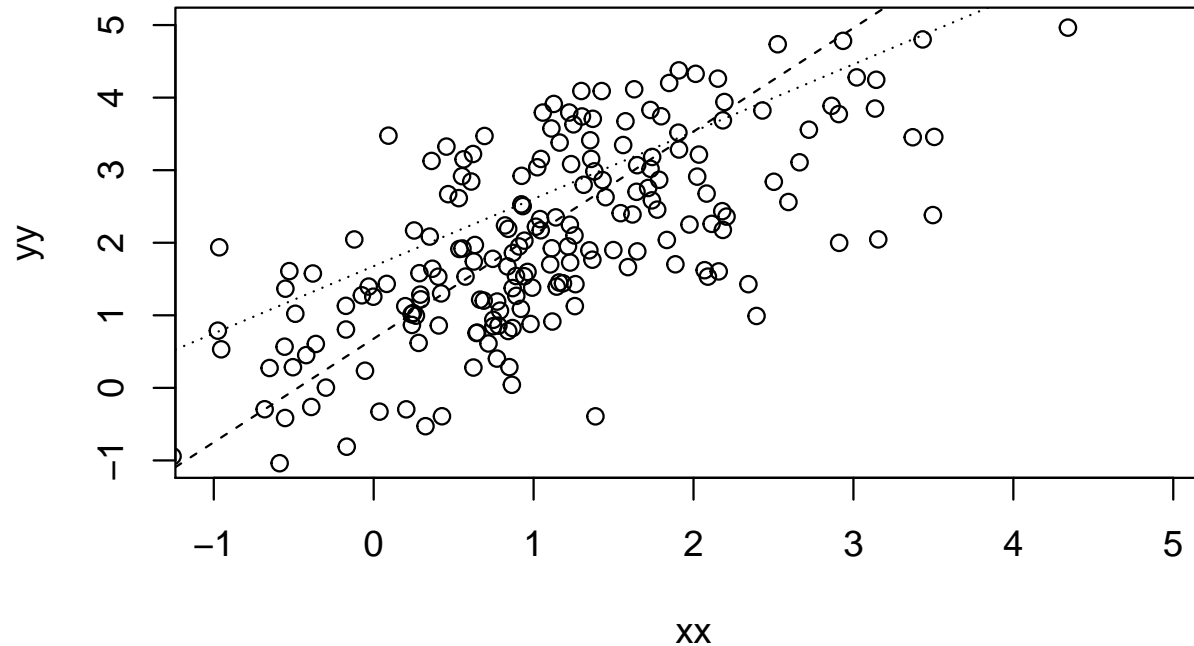
How we interpret the parameters β_0 , β_1 , and σ^2 in this model?



- β_0 : intercept, $E[y_i | x_i = 0]$
- β_1 : slope, change in $E[y_i | x_i]$ when $x_i \uparrow$ by 1 unit
- σ^2 : Variance of y_i about $E[y_i | x_i]$



Lets say that we knew β_0 , β_1 , and σ^2 , and had covariates x_1, \dots, x_n . Then, we could generate new data y_1, \dots, y_n using eq. 1.



However, in reality we typically only have observed data points $(x_1, y_1), \dots, (x_n, y_n)$ and do not know β_0 , β_1 , and σ^2 , and do not observe $\varepsilon_1, \dots, \varepsilon_n$. We estimate these quantities given the observed data. How to choose the best line?

Going back to the original example, we have the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, 50,$$

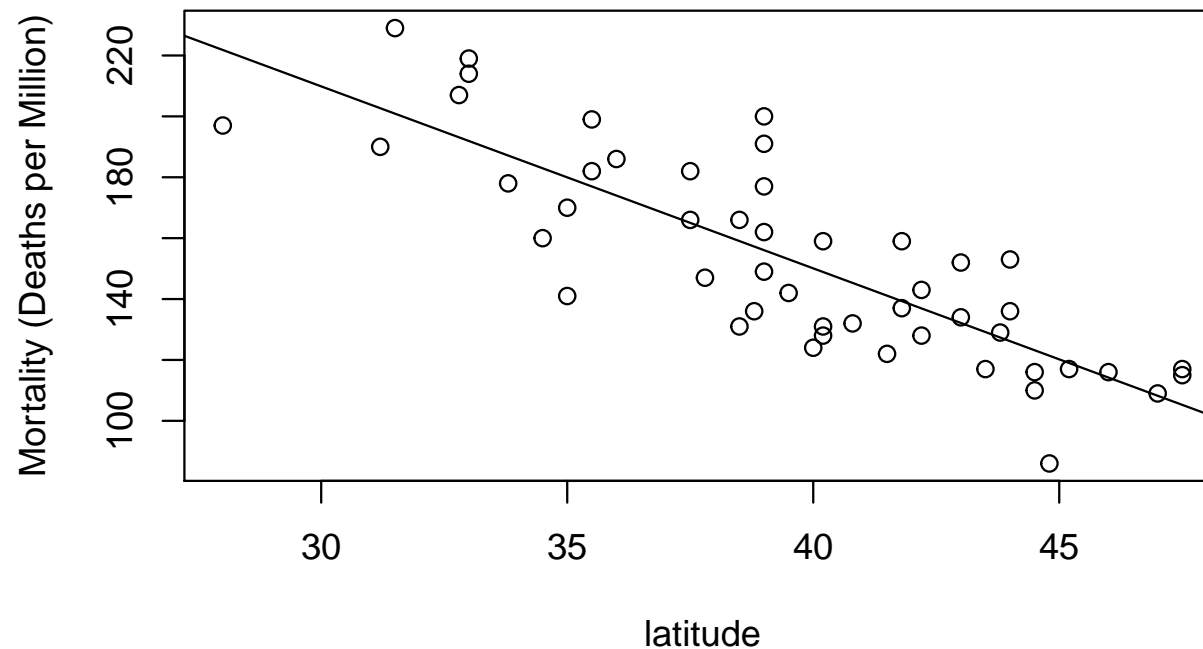
where (β_0, β_1) are *parameters* to be estimated. The ε_i 's represent random errors, which account for the fact that the response will vary even for states with the same latitudes (see, for example, MS and AL).

We may wish to use this model to do the following.

- Estimate β_0 , β_1 , and σ^2 .
- Test hypotheses about β_1 or obtain confidence limits for β_1 , relating the statistical results back to the scientific question concerning the relationship between latitude and melanoma mortality.
- Predict a future y at a given x .

Fitting the data, we obtain the estimates $\widehat{\beta}_0 = 389.189$ and $\widehat{\beta}_1 = -5.978$.

We can use these estimates to draw the estimated regression line.



How do we interpret our parameter estimates?

Ordinary Least Squares (OLS) Estimation

- Many possible values of β_0, β_1
- Many different possible lines
- OLS is an approach to obtain estimates of these parameters, $\hat{\beta}_0, \hat{\beta}_1$, in addition to $\hat{\sigma}^2$

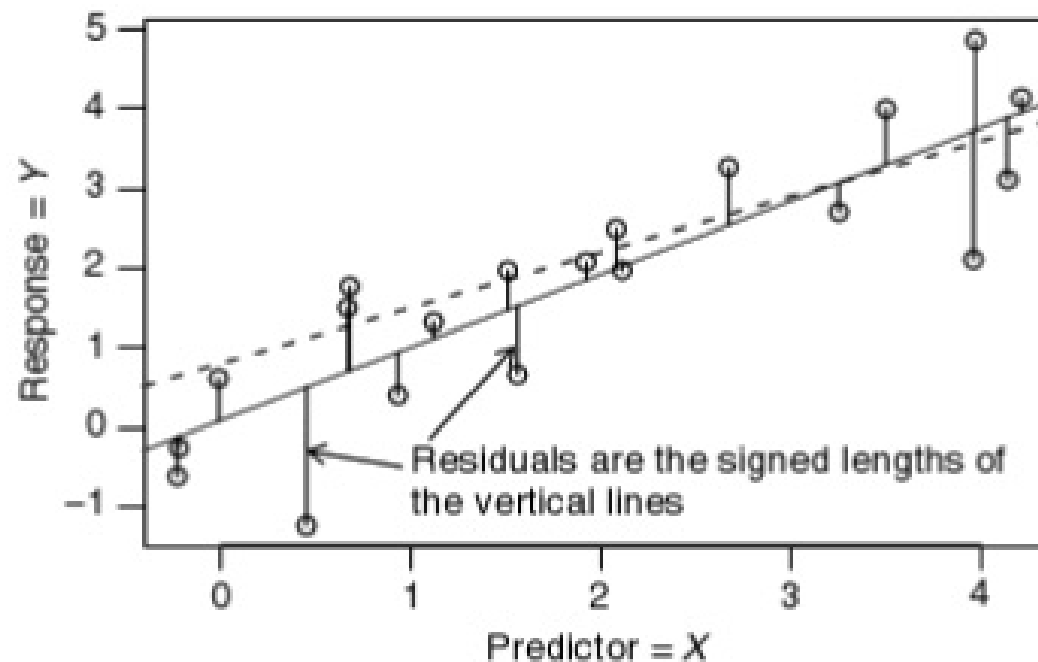
With these estimated values we may obtain two values

$$\hat{y}_i = \hat{E}[y_i|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \dots, n$$

where \hat{y}_i is the predicted value of y_i given $\hat{\beta}_0$, $\hat{\beta}_1$, x_i , and

$$\hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), i = 1, \dots, n$$

where $\hat{\varepsilon}_i$ is called the residual for the i th observation.



Intuition behind OLS - graphical illustration (Weisberg pg. 25). Solid line is the fitted line, dashed line is the true line (why are these different?) that the data was generated from. But how do we select the fitted line, and how do we know it is "best" given the observed data?

Least Squares Criterion The values for $(\hat{\beta}_0, \hat{\beta}_1)$ are obtained by minimizing the RSS with respect to (β_0, β_1) . That is, we minimize

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Define the Residual Sum of Squares (RSS), also known as the Sums of Squares for Error (SSE), as the following

$$\begin{aligned} RSS(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \\ &= S_{YY} - \hat{\beta}_1^2 S_{XX} \end{aligned}$$

where $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$ and $S_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2$

But how do we minimize this $RSS(\beta_0, \beta_1)$? Discussed in Weisberg A.3, we will briefly go over approach here.



Intuitively, $\hat{\sigma}^2$ can be estimated by averaging $\hat{\varepsilon}_i^2$. An unbiased estimate for $\hat{\sigma}^2$ is given by

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

This quantity is also referred to as the Residual Mean Square, or the Mean Square of the Errors (MSE). RSS may also be expressed as

$$RSS = \sum_{i=1}^n [\hat{\varepsilon}_i]^2$$

Why is this the case? Note that RSS is determined solely on $\hat{\beta}_0$ and $\hat{\beta}_1$.

Assuming that the errors are drawn from a normal distribution, then the Residual Mean Square divided by σ^2 will be distributed Chi-squared with $n - 2$ degrees of freedom. Then, we have that

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n - 2} \chi^2(n - 2)$$

so that

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{\sigma^2}{n - 2} E[\chi^2(n - 2)] \\ &= \frac{\sigma^2}{n - 2} (n - 2) \\ &= \sigma^2 \end{aligned}$$

→ $\hat{\sigma}^2$ is unbiased. This is in contrast to the MLE of σ^2 . 

Properties of Least Squares Estimates We can see that $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the random ε_i 's (why?). If the errors are mean 0, then

$$E[\hat{\beta}_0] = \beta_0$$

$$E[\hat{\beta}_1] = \beta_1$$



In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Notice that we do not need the assumption of normality for this result to hold. The normal assumption of the errors impacts hypothesis testing.

$$\begin{aligned} E[\hat{\beta}_0] &= \beta_0 \\ E[\hat{\beta}_1] &= \beta_1 \end{aligned}$$

Can you show how these estimators are unbiased?

Now lets look at the variances of teh estimates

$$Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \quad \text{🗨️}$$

$$Var[\hat{\beta}_1] = \frac{\sigma^2}{S_{XX}}$$

How do we obtain this result?

We can also derive the forms of the covariances and correlation of the estimates

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{S_{XX}}$$

$$\rho(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sqrt{\frac{S_{XX}}{n} + \bar{x}^2}}$$



What do these results imply? How do we obtain this result?

The Gauss-Markov theorem provides an optimality result for ols estimates. Among all estimates that are linear combinations of the y_1, \dots, y_n and unbiased, the ols estimates have the smallest variance. These estimates are called the best linear unbiased estimates, or blue. If one believes the model assumptions and is interested in using linear unbiased estimates, the ols estimates are the ones to use.

The means and variances, and covariances of the estimated regression coefficients do not require a distributional assumption concerning the errors. Since the estimates are linear combinations of the y_1, \dots, y_n , and hence linear combinations of the errors, the central limit theorem shows that the coefficient estimates will be approximately normally distributed if the sample size is large enough.

For smaller samples, if the errors are i.i.d, then the regression estimates $(\hat{\beta}_0, \hat{\beta}_1)$ will have a joint normal distribution with means, variances, and covariances as given before. When the errors are normally distributed, the OLS estimates can be justified using a completely different argument, since they are then also maximum likelihood estimates.

Estimated Variances

$$\widehat{Var}[\hat{\beta}_0] = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

$$\widehat{Var}[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{S_{XX}}$$

$$se(\hat{\beta}_j) = \sqrt{\widehat{Var}[\hat{\beta}_j]}, j = 1, 2$$

Standard deviation - variability of an observable random variable like the response y_i or an unobservable random variance like the errors ε_i .


Standard error - square root of the **estimated** variance of a statistic like $\hat{\beta}_1$.

Confidence Intervals and t-tests Confidence intervals result in interval estimates, while tests provide methodology for making decisions concerning the value of a parameter or fitted value.

When the errors are $\overset{i.i.d}{\sim} N(0, \sigma^2)$, parameter estimates, fitted values, and predictions will be normally distributed. Why is this the case?

Confidence intervals and tests can be based on a t-distribution, which is the appropriate distribution with normal estimates but using $\hat{\sigma}^2$ to estimate the unknown variance σ^2 . Suppose we let $t(\alpha/2, d)$ be the value that cuts off $\alpha/2 \times 100\%$ in the upper tail of the t-distribution with d df.

For example, the standard error of the intercept is

 $se(\hat{\beta}_0) = \hat{\sigma}(1/n + \bar{x}^2/S_{XX})$. Hence, a $(1 - \alpha) \times 100\%$ CI is the set of points β_0 in the interval

$$\hat{\beta}_0 - t(\alpha/2, n - 2)se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t(\alpha/2, n - 2)se(\hat{\beta}_0).$$

$1 - \alpha$ percent of such intervals will include the true value.

A hypothesis test of

$$H_0 : \beta_0 = \beta_0^*, \beta_1 \text{ arbitrary},$$

$$H_A : \beta_0 \neq \beta_0^*, \beta_1 \text{ arbitrary}$$


is obtained by computing the t-statistic $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}$ and referring this ratio to the t-distribution with $df = n - 2$, the number of df in the estimate of σ^2 .

From our Melanoma example, testing

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

would be addressing what question in particular?

Prediction If we found a new state and latitude value, x_{51} , can we predict what the mortality would be, given the fitted model? That is, we would like to know what y_{51} would be given x_{51} , assuming the existing data is relevant to the new data. A point prediction of y_{51} , say \tilde{y}_{51} , is simply 

$$\tilde{y}_{51} = \hat{\beta}_0 + \hat{\beta}_1 x_{51}$$

where \tilde{y}_{51} predicts the unobserved y_{51} .



Assuming the model is correct, then the true value of y_{51} is

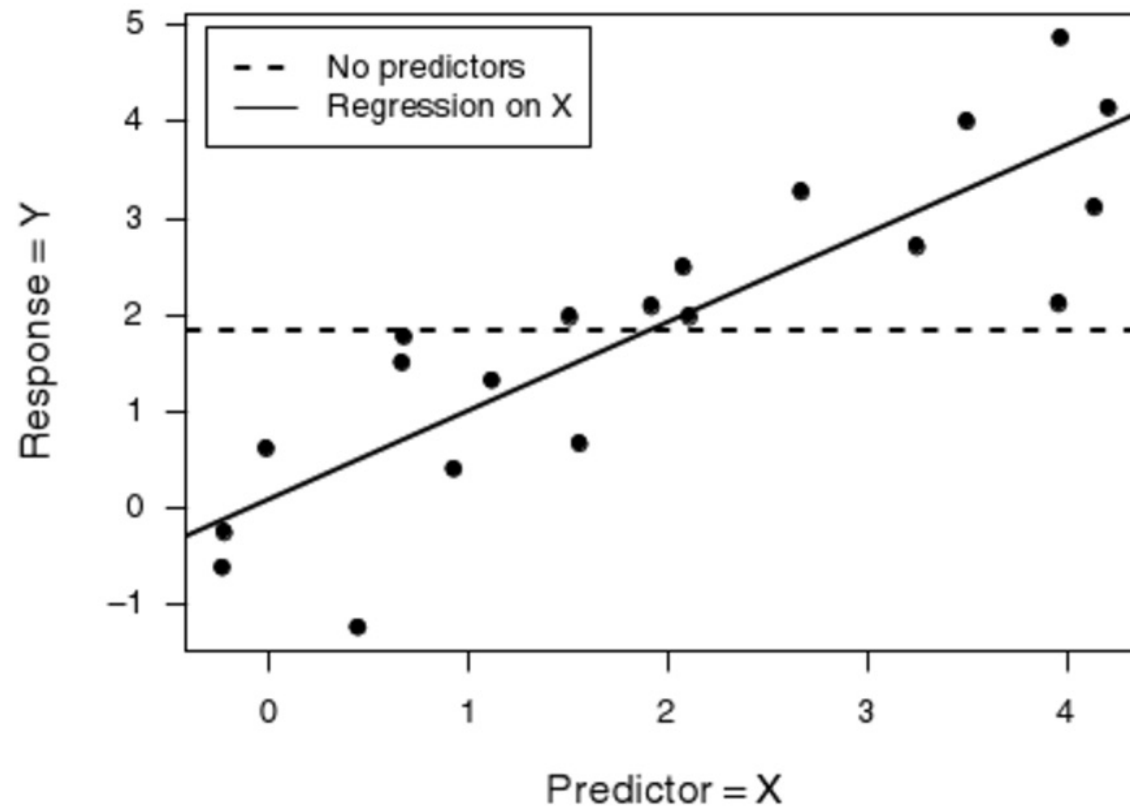
$$\tilde{y}_{51} = \beta_0 + \beta_1 x_{51} + \varepsilon_{51}$$

where ε_{51} is the random error for attached to the future value. Since \tilde{y}_{51} is based on estimated values of β_0 and β_1 , the prediction error variability will have a second component tha arises from the uncertainty in the estimtates of the coefficients. We can show that

$$Var(\tilde{y}_{51}|x_{51}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_{51} - \bar{x})^2}{S_{XX}} \right).$$

What does this imply?

Coefficient of Determination R^2

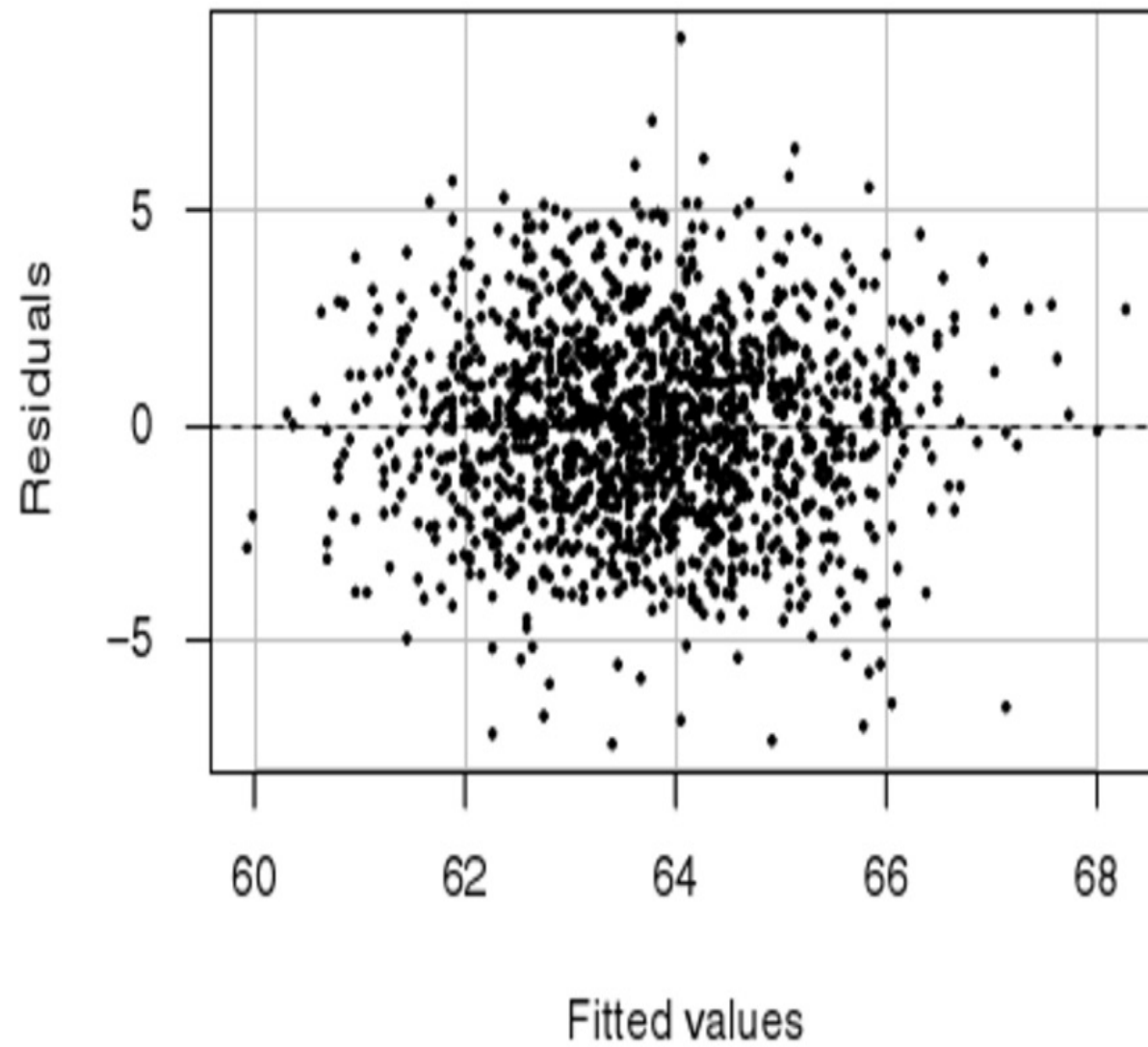


Ignoring all covariates, the best prediction of the response would be the total sum of squares $SSY = \sum_{i=1}^n (y_i - \bar{y})^2$, the observed total variation of the response, ignoring any and all predictors.

The unexplained variation is given by RSS , the sum of squared deviations from the fitted line. We define $SSreg = SSY - RSS$. If both sides are divided by S_{YY} , then we have $\frac{SSreg}{S_{YY}} = 1 - \frac{RSS}{S_{YY}} = R^2$. R^2 , the coefficient of determination, may be thought of the proportion of the total variability explained by the model, or 1 - the unexplained variability. This is a scale-free one-number summary of the strength of the relationship between x_i and y_i . In later lectures we will go more into this summary.

Residuals

Plots of residuals versus other quantities are used to find failures of assumptions. The most common plot, especially useful in simple regression, is the plot of residuals versus the fitted values. A null plot would indicate no failure of assumptions. Curvature might indicate that the fitted mean function is inappropriate. Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance. A few relatively large residuals may be indicative of outliers, cases for which the model is somehow inappropriate.



In a later lecture, we will learn how to use residuals to check model assumptions.

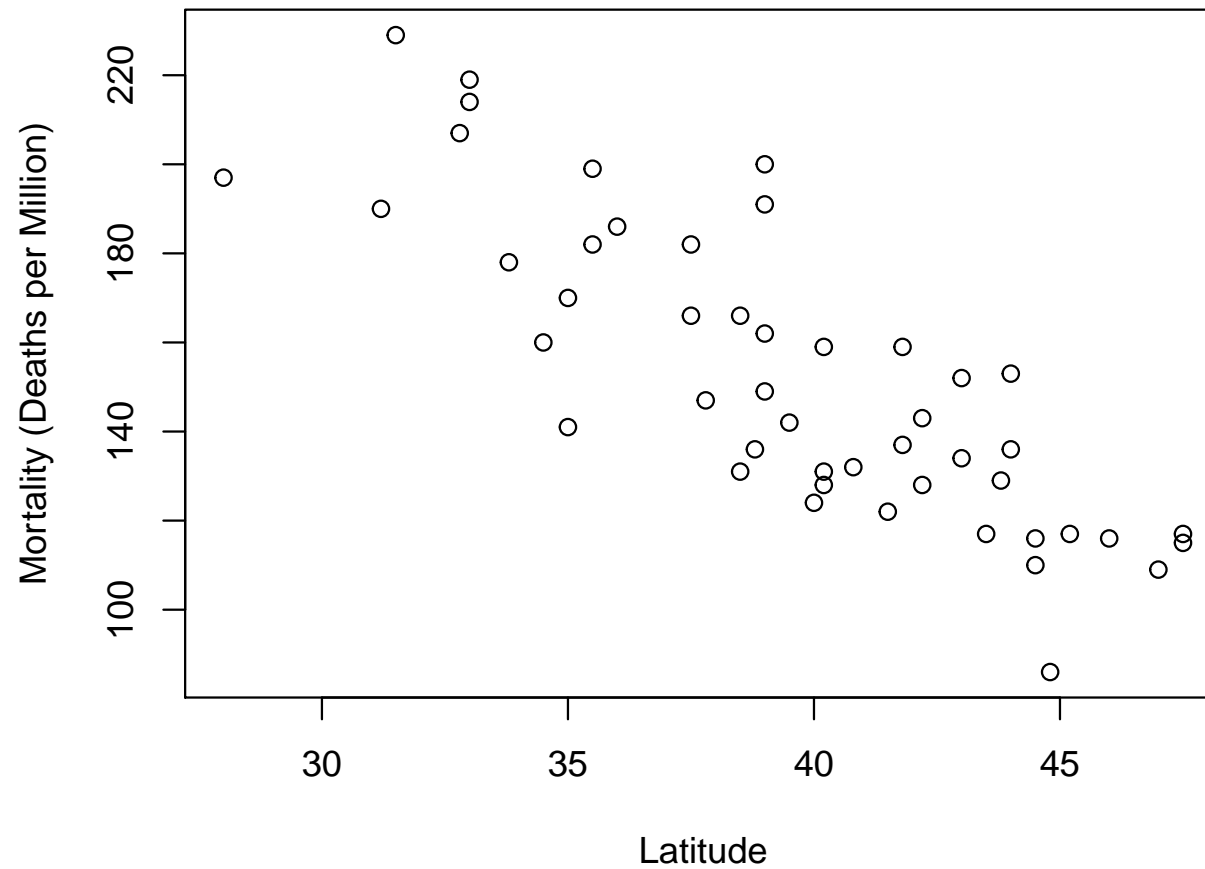
SLR in R and SAS

Lets take a peek at the data for the contiguous US

	mortality	latitude	longitude	ocean
Alabama	219	33.0	87.0	yes
Arizona	160	34.5	112.0	no
Arkansas	170	35.0	92.5	no
California	182	37.5	119.5	yes
Colorado	149	39.0	105.5	no
Connecticut	159	41.8	72.8	yes

Lets plot the data

```
> plot(USmelanoma[,2], USmelanoma[,1],  
+      ylab = "Mortality (Deaths per Million)",  
+      xlab = "Latitude"  
+ )
```



We can fit the linear regression model with the `lm` function

```
> out = lm(mortality ~ latitude, data = USmelanoma)
> out = lm(USmelanoma[,1] ~ USmelanoma[,2])
```

The summary function prints basic information about the fit

```
> summary(out)
```

Call:

```
lm(formula = USmelanoma[, 1] ~ USmelanoma[, 2])
```

Residuals:

Min	1Q	Median	3Q	Max
-38.485	-12.823	1.272	12.192	44.381

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	388.312	23.767	16.338	< 2e-16 ***
USmelanoma[, 2]	-5.966	0.597	-9.994	4.15e-13 ***

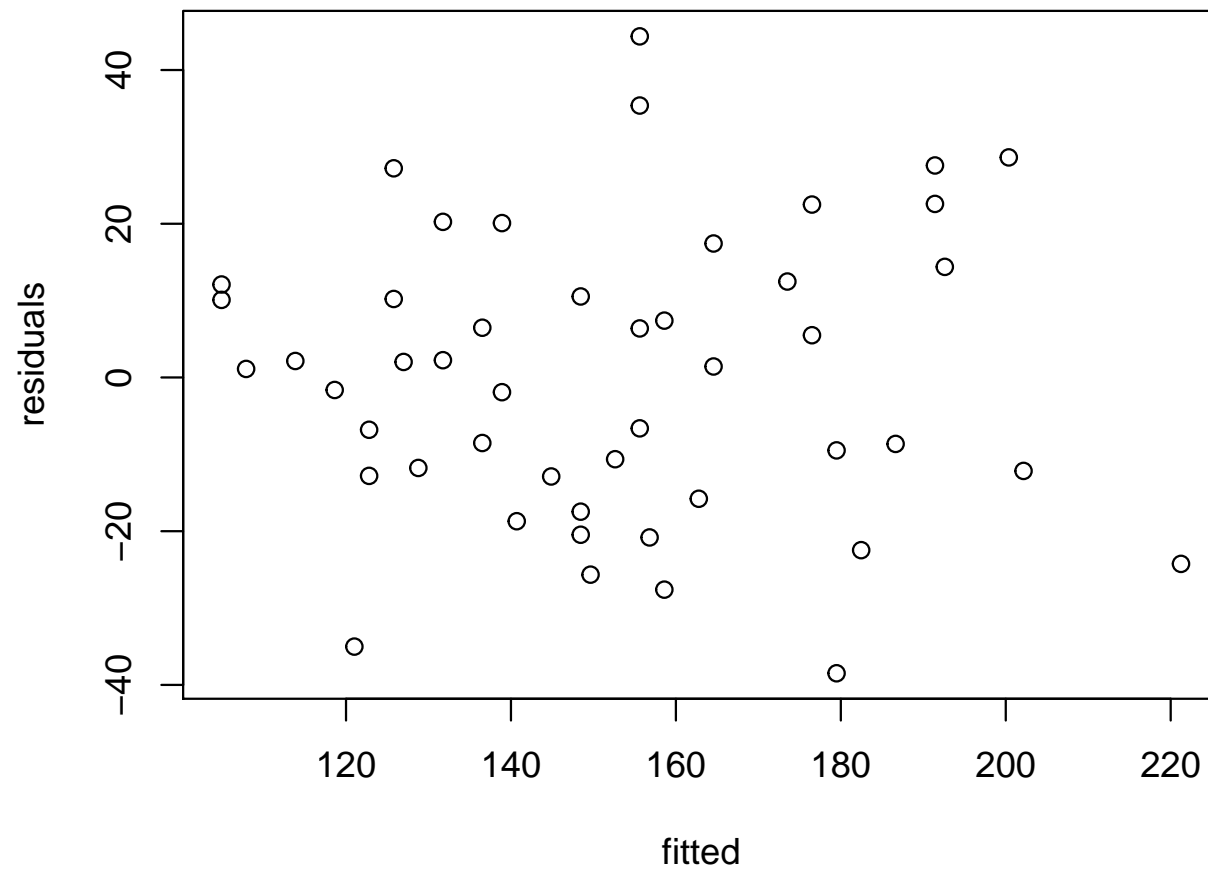
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.07 on 46 degrees of freedom
Multiple R-squared: 0.6847, Adjusted R-squared: 0.6778
F-statistic: 99.89 on 1 and 46 DF, p-value: 4.145e-13

We can also extract values based on the model fit

```
> residuals = out$residuals  
> fitted = out$fitted.values  
> coefficients = out$coefficients
```

```
> plot(fitted, residuals)
```



Now, lets use what we learned so far to check the results

```
> y = USmelanoma$mortality
> x = USmelanoma$latitude
> n = length(y)
> ybar = mean(y)
> xbar = mean(x)
> S_{YY} = sum( (y - ybar)^2 )
> S_{XX} = sum( (x - xbar)^2 )
> S_{XY} = sum( (y - ybar)*(x - xbar) )
```

Now, lets use what we learned so far to check the results

```
> coefficients
```

```
      (Intercept) USmelanoma[, 2]  
      388.311830      -5.966476
```

```
> S_{XY}/S_{XX} # beta_1 hat
```

```
[1] -5.966476
```

```
> ybar - (S_{XY}/S_{XX})*xbar # beta_0 hat
```

```
[1] 388.3118
```

```

> RSS = sum( (y - fitted)^2 )
> s2_hat = RSS/(n - 2) # sigma_squared hat
> s2_hat

[1] 363.5953

> sqrt(s2_hat*(1/n + xbar^2/S_{XX})) # se of beta_0 hat

[1] 23.76711

> sqrt(s2_hat/S_{XX}) # se of beta_1 hat

[1] 0.5969898

> round(vcov(out),3) # covariance matrix of beta_0 hat, beta_1 hat

              (Intercept) USmelanoma[, 2]
(Intercept)      564.875      -14.093
USmelanoma[, 2]  -14.093       0.356

```

From our Melanoma example, testing

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

$$t = \frac{-5.966 - 0}{0.597} = -9.994$$

```
> t = -9.994  
> p.value = 2*(1 - pt(abs(t), n-2))  
> p.value  
  
[1] 4.147793e-13
```

SAS Proc Reg

```
proc reg data = USmelanoma;  
    model mortality = latitude;  
run;
```

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: mortality

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	36464	36464	99.80	<.0001
Error	47	17173	365.38436		
Corrected Total	48	53637			

Root MSE	19.11503	R-Square	0.6798
Dependent Mean	152.87755	Adj R-Sq	0.6730
Coeff Var	12.50349		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	389.18935	23.81232	16.34	<.0001
latitude	1	-5.97764	0.59837	-9.99	<.0001

Mini Problems to do from Weisberg

- 2.1.3
- 2.11