Lecture 3
Simple linear regression, cont.

BIOST 515

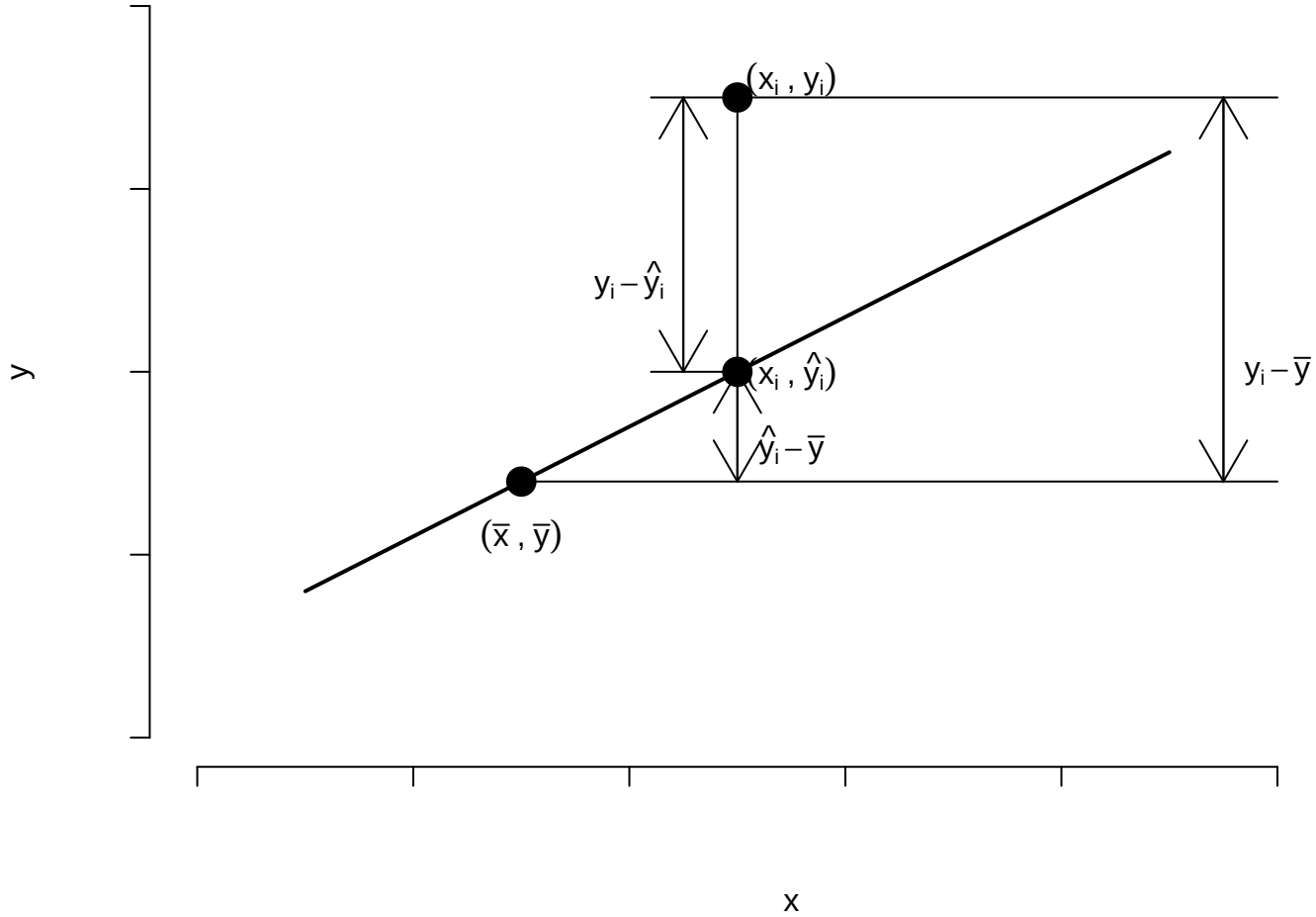
January 13, 2004

Breakdown of sums of squares

The simplest regression estimate for Y_i is $\bar{Y}$ (an intercept-only model). $Y_i - \bar{Y}$ is the total error and can be broken down further by

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

total error = residual error + error explained by regression


(x_i , y_i)
$y_i - \hat{y}_i$
$(x_i , \hat{y}_i)$
$\hat{y}_i - \bar{y}$
$y_i - \bar{y}$
$(\bar{x} , \bar{y})$
y
x

If we square the previous expression and sum over all observations, we get

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

$=$

$$SSTO = SSR + SSE,$$

where $SSTO$ is the corrected sums of squares of the observations, SSR is the sum of squares regression and SSE is the sums of squares error.

Intuitively, if SSR is 'large' compared to SSE, then β_1 is significantly different than zero.

Recall that $Z_2 = \frac{SSE}{\sigma^2} \sim \chi^2_{N-2}$. It can also be shown that, under H_0, $Z_1 = \frac{SSR}{\sigma^2}\ \chi^2_1$ and Z_1 and Z_2 are independent. Under H_0,

$$F = \frac{Z_1/1}{Z_2/(N-2)} = \frac{SSR}{SSE/(N-2)} \sim F_{1,N-2}.$$

If the observed statistic

$$F_{obs} > F_{1,N-2,1-\alpha},$$

then we reject $H_0 : \beta_1 = 0$.

The calculations for the F-test are usually presented in an analysis of variance (ANOVA) table.

Source of variation	Sums of squares	Degrees of freedom	Mean square	E[Mean square]
Regression	$SSR = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$	1	SSR	$\sigma^2 + \beta_1^2 \sum_{i=1}^{N}(X_i - \bar{X})^2$
Error	$SSE = \sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2$	N-2	$\frac{SSE}{N-2}$	σ^2
Total	$SSTO = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$	N-1		

```
lm1=lm(Mortality~Education,data=smsa)
anova(lm1)

Analysis of Variance Table

Response: Mortality
          Df Sum Sq Mean Sq F value    Pr(>F)
Education  1  59662   59662  20.508 3.008e-05 ***
Residuals 58 168737    2909
```

$F_{obs} = 59662/(168737/58) = 20.51 > F_{1,58,.95} = 4.01.$
Therefore, we reject $H_0 : \beta_1 = 0$.
To get SSTO:

```
alm1=anova(lm1)
SSTO=sum(alm1$"Sum Sq")
print(SSTO)
```

```
[1] 228398.3
```

Where do the degrees of freedom come from?

In class, we will show that the t-test and F-test are equivalent for $H_0 : \beta_1 = 0$. However, the t-test is somewhat more adaptable as it can be used for one-sided alternatives. We can also easily calculate it for different hypothesized values in H_0.
One-sided t-test for the SMSA example:
$H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 < 0$.

$$t_{obs} = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = -4.529$$

$t_\alpha^{N-2} = -1.627 > -4.529$ therefore reject H_0 in favor of H_A.

Coefficient of Determination

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

- Often referred to as the proportion of variation explained be the predictor
- Because $0 \leq SSE \leq SSTO$, $0 \leq R^2 \leq 1$
- As predictors are added to the model R^2 will not decrease
- Large R^2 does not necessarily imply a "good" model

- R^2 does not
  - ⋆ measure the magnitude of the slope
  - ⋆ measure the appropriateness of the model

From SMSA example with education as a predictor of mortality:

```
R2=alm1$"Sum Sq"[1]/SSTO
print(R2)
```

```
0.261217
```

$R^2 = 0.26$

Prediction

Sometimes, we would like to be able to predict the outcome for a new value of the predictor. The new outcome is defined as

$$y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon$$

with an estimated value of

$$\widehat{y_{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{new} + \hat{\epsilon}.$$

The expected value is

$$E[\widehat{y_{new}}] = \beta_0 + \beta_1 x_{new},$$

and the variance is

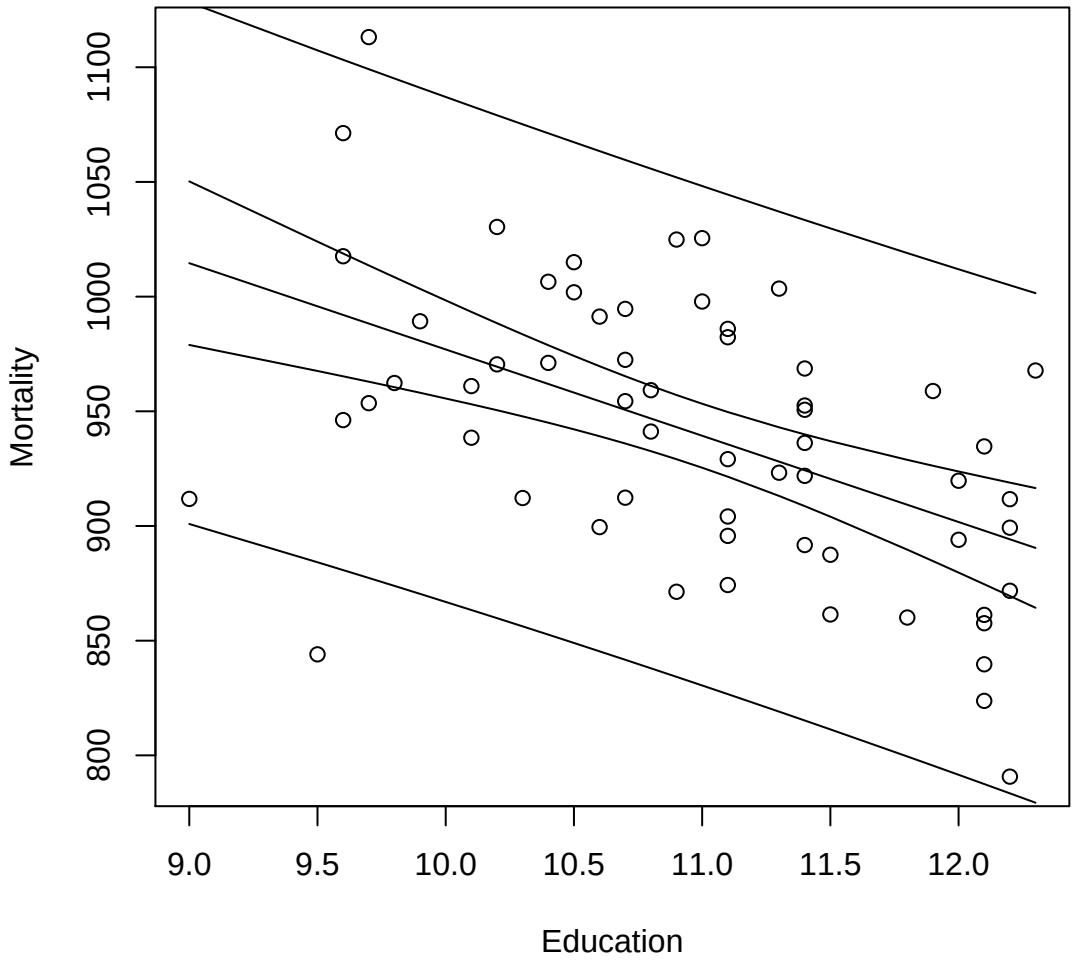
$$\text{var}(\widehat{y_{new}}) = \sigma^2 \left\{ 1 + \frac{1}{N} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \right\}.$$

The $100(1-\alpha)\%$ confidence interval is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} \pm t_{N-2,1-\alpha/2} \times \hat{\sigma} \times \left\{ 1 + \frac{1}{N} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \right\}^{1/2}.$$

Note: We have assumed $\epsilon \sim N(0, \sigma^2)$ to construct the prediction interval. If the error terms are not close to normal,

then the prediction interval could be misleading. This is not the case for the interval for the fitted response which only requires approximate normality for $\hat{\beta}_0$ and $\hat{\beta}_1$.


Mortality
800
850
900
950
1000
1050
1100
9.0
9.5
10.0
10.5
11.0
11.5
12.0
Education

Maximum Likelihood Estimation

Assumptions about the distribution of ϵ_i are not necessary for least squares estimation. If we assume that $\epsilon_i \sim_{iid} N(0, \sigma^2)$, then $Y_i \sim_{iid} N(\beta_0 + \beta_1 x_i, \sigma^2)$ and

$$p(Y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 x_i)^2\}.$$

The likelihood is then equal to

$$L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1 x_i)^2\}.$$

The maximum likelihood estimators (MLEs) are those values of β_0, β_1 and σ^2 that maximize L or, equivalently, $l = log(L)$.

$$l \propto -N/2 \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

The MLEs for the simple linear regression model are given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{(x)},$$

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

and

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

The MLEs for β_0 and β_1 are the same as the least squares estimators. However the MLE for σ^2 is not. Recall that the least squares estimate of σ^2 is unbiased. The MLE of σ^2 is biased (although it is consistent).

Considerations in the use of regression

1. Regression models are only interpretable over the range of the observed data.

2. The disposition of x plays an important role in the model fit.

3. Outliers or erroneous data can disturb the model fit.

4. Just because the regression results indicate that two variables are related, there is no evidence about causality.

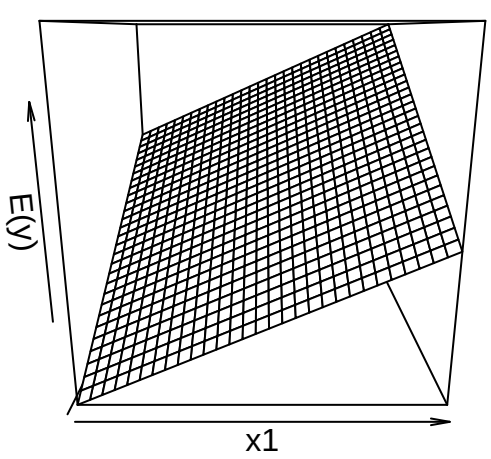
Multiple Linear Regression

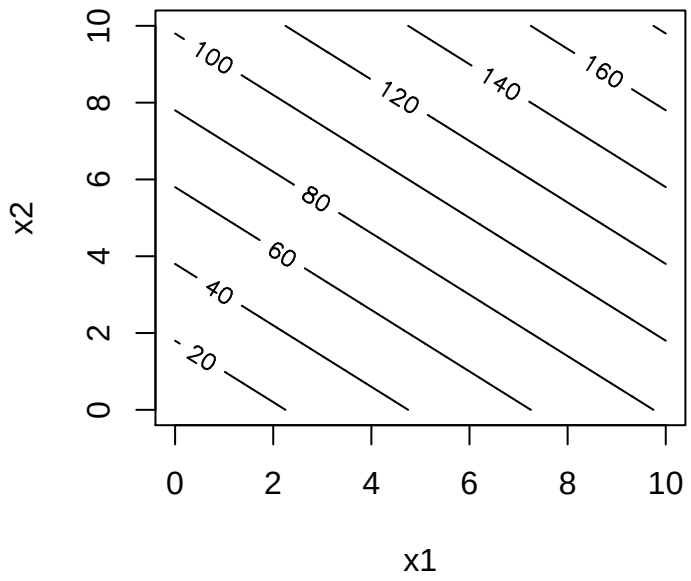
Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

$$E(y) = 2 + 8x_1 + 10x_2$$

β_1 indicates the change in the expected response per unit change in x_1 when x_2 is held constant. Likewise, β_2 represents the change in the expected response per unit change in x_2 when x_1 is held constant.


E(y)
x1


x2
0
2
4
6
8
10
20
40
60
80
100
120
140
160
x1

We now consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

$i = 1, \ldots, n$, $E[\epsilon_i] = 0$, var$(\epsilon_i) = \sigma^2$ and $cov(\epsilon_i, \epsilon_j) = 0$. The parameter $\beta_j, j = 1 \ldots, p$ represents the expected change in y_i per unit of change in x_j holding the remaining predictors $x_i (i \neq j)$ constant.

We can use the model defined in (1) to describe more complicated models. For example, we might be interested in a cubic polynomial model,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon.$$

If we let $x_1 = x$, $x_2 = x^2$ and $x_3 = x^3$, then we can rewrite the regression model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

which is a multiple linear regression model with 3 predictors. How do we interpret this model?

Interactions

We may also want to include *interaction effects*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

If we let $x_3 = x_1 x_2$, this model is equivalent to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$