ST430: Introduction to Regression Analysis, Ch3, Sec 3.1-3.6

Luo Xiao

August 31, 2015
Simple Linear Regression
Recall:

A regression model describes how a *dependent* variable (or *response*) $Y$ is affected, on average, by one or more *independent* variables (or *factors*, or *covariates*) $X_1, X_2, \ldots, X_p$. 
The straight-line probabilistic model

Simplest case of a regression model:

- One independent variable, \( p = 1, X_1 \equiv X; \)
- Linear dependence;
- Model equation:

\[
E(Y) = \beta_0 + \beta_1 X,
\]

or equivalently

\[
Y = \beta_0 + \beta_1 X + \epsilon.
\]
**Interpreting the parameters:**

- $\beta_0$ is the *intercept* (so called because it is where the graph of $y = \beta_0 + \beta_1 x$ meets the $y$-axis $x = 0$);

- $\beta_1$ is the *slope*; that is, the change in $E(Y)$ as $X = x$ is changed to $X = x + 1$.

Note: if $\beta_1 = 0$, $X$ has no effect on $y$; that will often be an interesting *hypothesis* to test.
Graph for a straight line

\[ E(y) = \beta_0 + \beta_1 x \]

\[ \beta_0 = y \text{- intercept} \]

\[ \beta_1 = \text{Slope} \]
Advertising and Sales example

- $x = \text{monthly advertising expenditure, in hundreds of dollars;}$
- $y = \text{monthly sales revenue, in thousands of dollars;}$
- $\beta_0 = \text{expected revenue with no advertising;}$
- $\beta_1 = \text{expected revenue increase per$100$increase in advertising, in thousands of dollars.}$

Sample data for five months:

<table>
<thead>
<tr>
<th>Advertising</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revenue</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Scatterplot

R code:

```r
x = c(1,2,3,4,5)
y = c(1,1,2,2,4)
plot(x,y,type="p",pch=19,
   xlim=c(0,5),ylim=c(0,4),
   xlab="Advertising expenditure",ylab="Sales revenue")
```
R output

Simple Linear Regression

Sales revenue vs Advertising expenditure

- R output
- Simple Linear Regression
We could try various values of $\beta_0$ and $\beta_1$.

For given values of $\beta_0$ and $\beta_1$, we get *predictions*

$$p_i = \beta_0 + \beta_1 x_i, \ i = 1, 2, 3, 4, 5.$$  

The difference between the observed value $y_i$ and the prediction $p_i$ is the *residual*

$$r_i = y_i - p_i, \ i = 1, 2, 3, 4, 5.$$  

A good choice of $\beta_0$ and $\beta_1$ gives accurate predictions, and generally small residuals.
One candidate line ($\beta_0 = -0.1, \beta_1 = 0.7$):

R code

```r
x = c(1,2,3,4,5)
y = c(1,1,2,2,4)
plot(x,y,type="p",pch=19,
    xlim=c(0,5),ylim=c(0,4),
    xlab="Advertising expenditure",ylab="Sales revenue")
abline(a=-0.1,b=0.7,col="red")
```

Simple Linear Regression
R output

Simple Linear Regression
Fitting the model

How to measure the overall size of the residuals?

Most common measure (but not the only possibility): sum of squares of residuals

\[
\sum r_i^2 = \sum (y_i - p_i)^2 \\
= \sum \{y_i - (\beta_0 + \beta_1 x_i)\}^2 \\
= S(\beta_0, \beta_1).
\]

The least squares line is the one with the smallest sum of squares.

Note: the least squares line has the property that \( \sum r_i = 0 \); Definition 3.1 (page 95) does not need to impose that as a constraint.
The *least squares estimates* of $\beta_0$ and $\beta_1$ are the coefficients of the least squares line.

Some algebra shows that the least squares estimates are

\[
\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}
\]

and

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.
\]
Other criteria

Why square the residuals?

We could use *least absolute deviations* estimates, minimizing

\[
S_1(\beta_0, \beta_1) = \sum |y_i - (\beta_0 + \beta_1 x_i)|.
\]

Convenience: we have equations for the least squares estimates, but to find the least absolute deviations estimates we have to solve a *linear programming* problem.

Optimality: least squares estimates are BLUE if the errors \(\epsilon\) are uncorrelated with constant variance, and MVUE if additionally \(\epsilon\) is normal.
Model assumptions

The least squares line gives point estimates of $\beta_0$ and $\beta_1$.

These estimates are always unbiased.

To use the other forms of statistical inference:
- interval estimates, such as confidence intervals;
- hypothesis tests;
we need some assumptions about the random errors $\epsilon$. 
Zero mean: $E(\epsilon_i) = 0$; as noted earlier, this is not really an assumption, but a consequence of the definition

$$\epsilon = Y - E(Y).$$

Constant variance: $V(\epsilon_i) = \sigma^2$; this is a nontrivial assumption, often violated in practice.

Normality: $\epsilon_i \sim N(0, \sigma^2)$; this is also a nontrivial assumption, always violated in practice, but sometimes a useful approximation.

Independence: $\epsilon_i$ and $\epsilon_j$ are statistically independent; another nontrivial assumption, often true in practice, but typically violated with time series and spatial data.
Model assumptions

\[ E(y) = \beta_0 + \beta_1 x \]
Notes:

Assumptions 2 and 4 are the conditions under which least squares estimates are BLUE (Best Linear Unbiased Estimators);

Assumptions 2, 3, and 4 are the conditions under which least squares estimates are MVUE (Minimum Variance Unbiased Estimators).
Estimating $\sigma^2$

Recall that $\sigma^2$ is the variance of $\epsilon_i$, which we have assumed to be the same for all $i$.

That is,

$$
\sigma^2 = V(\epsilon_i) = V[Y_i - E(Y_i)] = V[Y_i - (\beta_0 + \beta_1 X_i)], \ i = 1, 2, \ldots, n.
$$

We observe $Y_i = y_i$ and $X_i = x_i$; if we knew $\beta_0$ and $\beta_1$, we would estimate $\sigma^2$ by

$$
\frac{1}{n} \sum \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = \frac{1}{n} S(\beta_0, \beta_1).
$$
We do not know $\beta_0$ and $\beta_1$, but we have least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

So we could use $S(\hat{\beta}_0, \hat{\beta}_1)$ as an approximation to $S(\beta_0, \beta_1)$.

But we know that $S(\hat{\beta}_0, \hat{\beta}_1) < S(\beta_0, \beta_1)$, so

$$\frac{1}{n} S(\hat{\beta}_0, \hat{\beta}_1)$$

would be a biased estimate of $\sigma^2$. 

Simple Linear Regression
We can show that, under Assumptions 2 and 4,

\[ E\left[ S\left( \hat{\beta}_0, \hat{\beta}_1 \right) \right] = (n - 2)\sigma^2. \]

So

\[ s^2 = \frac{1}{n-2} S\left( \hat{\beta}_0, \hat{\beta}_1 \right) = \frac{1}{n-2} \sum(y_i - \hat{y}_i)^2, \]

where \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \), is an unbiased estimate of \( \sigma^2 \).

This is sometimes written

\[ s^2 = \text{Mean Square Error} = \text{MSE} = \frac{\text{SSE}}{\text{degrees of freedom for Error}}. \]
Inferences about the line

We are often interested in the question of whether $X$ has any effect on $E(Y)$.

Since

$$E(Y) = \beta_0 + \beta_1 X,$$

the independent variable $X$ has some effect whenever $\beta_1 \neq 0$.

So we need to test the null hypothesis $H_0 : \beta_1 = 0$. 
We also need to construct a *confidence interval* for $\beta_1$, to indicate how precisely we know its value.

For both purposes, we need the *standard error*:

$$
\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}},
$$

where

$$
SS_{xx} = \sum (x_i - \bar{x})^2.
$$

As always, since $\sigma$ is unknown, we replace it by its estimate $s$, to get the *estimated standard error*

$$
\hat{\sigma}_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}. 
$$
A confidence interval for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \times \hat{\sigma}_{\beta_1}.$$ 

Note that we use the $t$-distribution with $n - 2$ degrees of freedom, because that is the degrees of freedom associated with $s^2$.

To test $H_0: \beta_1 = 0$, we use the test statistic

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}},$$

and reject $H_0$ at the significance level $\alpha$ if

$$|t| > t_{\alpha/2, n-2}.$$
Compare the confidence interval and the hypothesis test

Note that we reject $H_0$ if and only if the corresponding confidence interval does not include 0.
Linear regression in R: Advertising and Sales example

\[
x = c(1, 2, 3, 4, 5) \\
y = c(1, 1, 2, 2, 4) \\
fit = lm(y \sim x) \\
summary(fit)
\]

Try yourself to get the R output!
Construct confidence interval in R:

Use R command: `confint(fit)`

R output

```r
## 2.5 %  97.5 %
## (Intercept) -2.12112485 1.921125
##     x         0.09060793 1.309392
```
ANOVA table

Use R command: `anova(fit)`

R output:

```r
## Analysis of Variance Table
## Response: y
## Df   Sum Sq Mean Sq F value  Pr(>F)
## x    1    4.9  4.9000 13.3640 0.03535 *
## Residuals 3  1.10  0.3667
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```