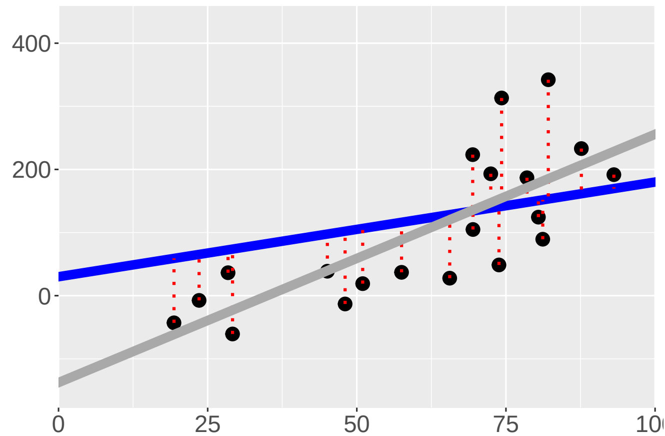
Chapter 2

The Simple Regression Model

Jim Bang

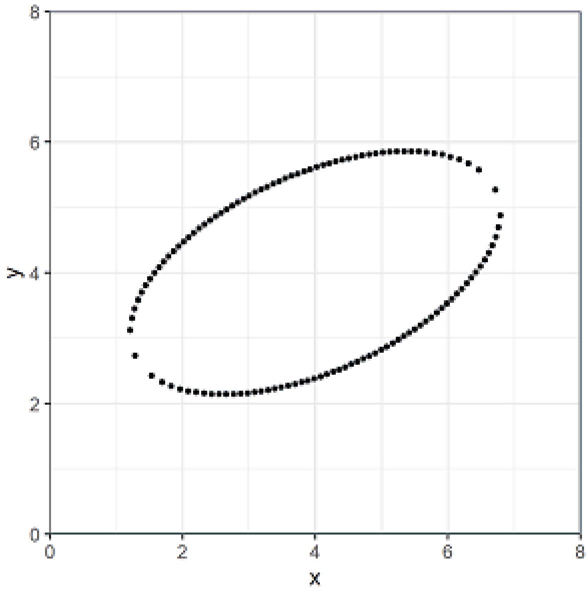
Ocular Estimation

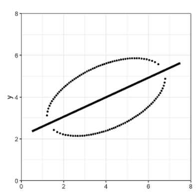
```
## [1] Intercept Guess: 30
## [1] Slope Guess: 1.5
## [1] MSE (Guess): 8655.7
## [1] True Intercept: -137.665
## [1] True Slope: 3.933
## [1] Mean squared error: 5473.8
```

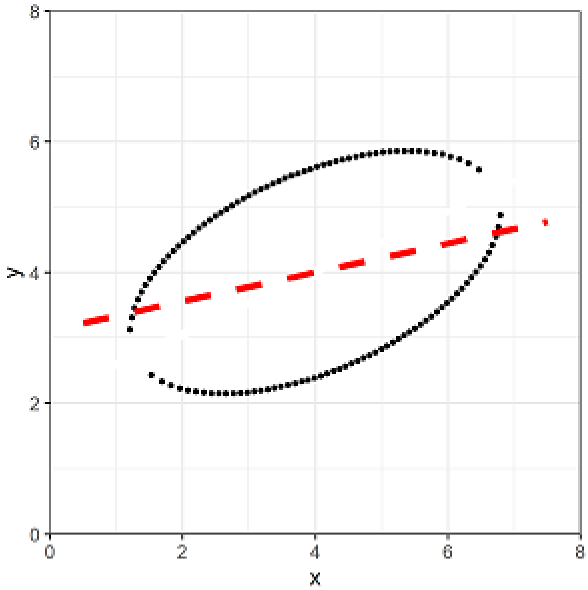


Regression Lines

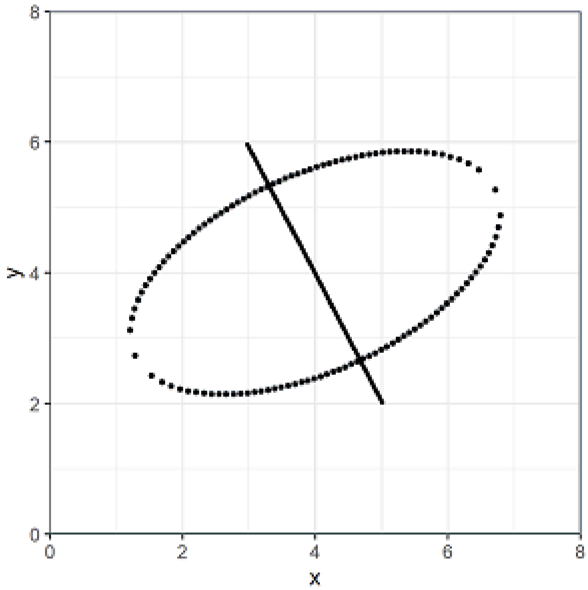
Which of the following graphs the best-fit regression linefor the following data?







8
6
4
2
0
y
0
2
4
6
8
x



8
6
4
2
0
y
0
2
4
6
8
x

A Simple Regression

Regression concepts

- Population Regression Function: $Wage = \beta_0 + \beta_1 Education + u$ β_0 and β_1 represent *parameters* - the true (but unknown) values for the intercept and the slope, respectively. u is the *error* - the true, random variation of wages that education doesn't capture. $Wage$ is the *endogenous* variable (or dependent variable/explained variable/predicted variable/response/regressand). $Education$ is assumed to be an *exogenous* variable (or independent variable/explanatory variable/predictor/control/regressor).
- Estimated Regression Line: $Wage = \hat{\beta}_0 + \hat{\beta}_1 Education + \hat{u}$ $\hat{\beta}_0$ and $\hat{\beta}_1$ represent *estimators* - the derived methods for estimating the intercept and the slope (e.g. OLS estimators). $\hat{u}$ is the *residual* - the observable deviations from the predicted wage and the actual wage for each observation. If we focus on the predictions, $\hat{wage} = \hat{\beta}_0 + \hat{\beta}_1 educ$ $\hat{wage}$ is the predicted wage

Assumptions

1. Errors have mean equal to zero, $E(u) = 0$.
2. Errors and X are independent, $E(u|X) = E(u) = 0$.

$E(u|X) = E(u) = 0 \Rightarrow Cov(u, X) = 0 \Rightarrow E(uX) = 0$

Click here to see the proof that $E(u|X) = 0 \Rightarrow E(ux) = 0$.

$$Cov(u, X) = E[(u - E(u))(X - E(X))] == E(uX) - E(u)E(X)$$

Since $E(u) = 0$, showing $Cov(u, X) = 0$ requires showing $E(uX) = 0$.

$$E_{u,X}(uX) = E_X(E_{u|X}(uX|X))$$

by the Law of Iterated Expectations (in reverse?).

$$E(E(uX|X)) = \int f_X(X) \int uX f_{u|X}(u|X) du dx = \int X f_X(X) \int u f_{u|X}(u|X) du dx = E_X(XE(u|X))$$
$$E_X(XE(u|X)) = E(X)E(u) = 0$$

by the Law of Iterated Expectations (again) and since $E(u) = 0$.

Note that even if $E(u) \neq 0$, $E(u|X) = 0 \Rightarrow Cov(u, X) = 0$ since $E(uX) = E(u)E(X)$.

$$E(y|x) = E(\beta_0 + \beta_1 x + u) = \beta_0 + \beta_1 E(x|x) + E(u|x)$$

The zero conditional mean condition guarantees that:

$$E(y|x) = \beta_0 + \beta_1 x.$$

This also guarantees that:

$$\Delta E(y|x) = \beta_1 \Delta x.$$

Wages and Education

1. Using the `wage1` data, regress wages on education (and an intercept). Name this `wage.lm1`
2. Summarize the regression object using `summary()`

```
wage.lm1 <- lm(wage ~ educ, data = wage1)
summary(wage.lm1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3396 -2.1501 -0.9674  1.1921 16.6085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90485    0.68497  -1.321    0.187
## educ         0.54136    0.05325  10.167   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 524 degrees of freedom
## Multiple R-squared:  0.1648, Adjusted R-squared:  0.1632
## F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

Regression without an Intercept

Duplicate this regression *without an intercept* and name it wage.lm0

```
wage.lm0 <- lm(wage ~ educ - 1, data = wage1)
summary(wage.lm0)
```

```
##
## Call:
## lm(formula = wage ~ educ - 1, data = wage1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.142 -2.246 -1.066  1.154 16.528
```

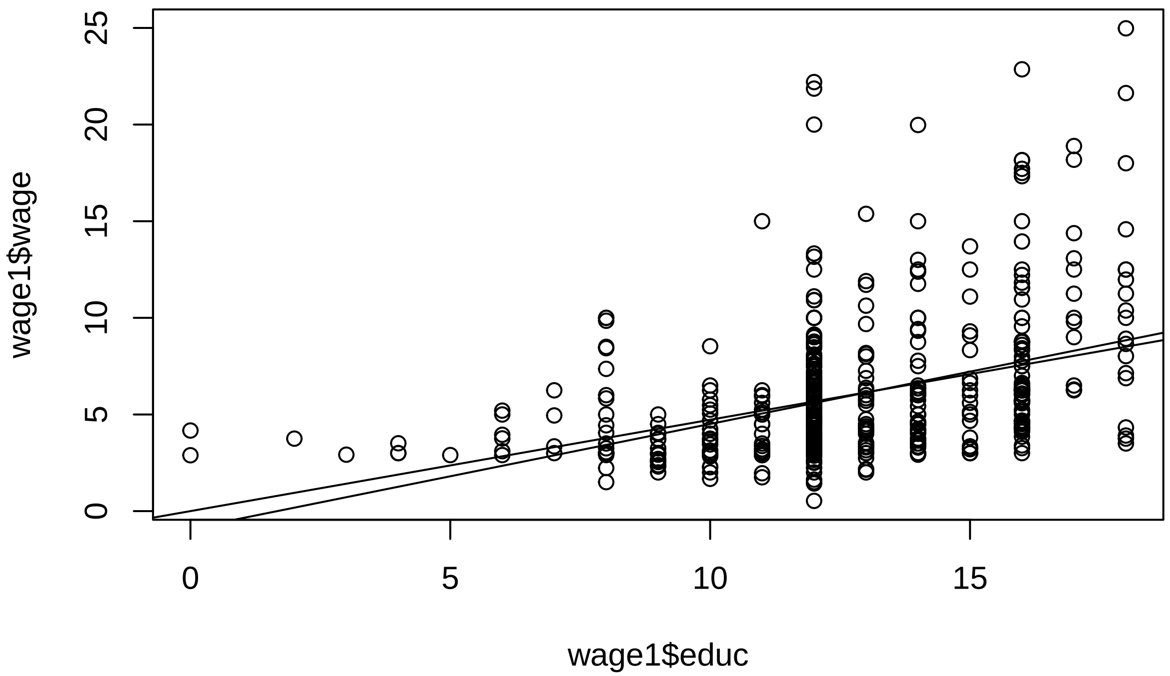
```
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## educ  0.47266    0.01146   41.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.381 on 525 degrees of freedom
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.7637
## F-statistic:  1701 on 1 and 525 DF,  p-value: < 2.2e-16
```

Plotting the Regression Line

Plot the following with base R graphics.

1. A scatter of the data for wage against education
2. The linear fit for the simple regression of wages on education *with* an intercept (dashed)
3. The linear fit for the simple regression of wages on education *without* an intercept (dotted)

```r
plot(wage1$educ, wage1$wage)
abline(wage.lm1)
abline(wage.lm0)
```


wage1$wage
wage1$educ
0
5
10
15
20
25
0
5
10
15

Properties of OLS

1. $\sum_{i=1}^{n} \hat{u}_i = 0$
2. $\sum_{i=1}^{n} x_i \hat{u}_i = 0$
3. The *Total* Sum of Squares, $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
4. The *Explained* Sum of Squares, $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
5. The *Residual* Sum of Squares, $SSR = \sum_{i=1}^{n} (y_i - \hat{y})^2$
6. SST = SSE + SSR (see book section 2.3 for proof)
7. Goodness of Fit, $R^2 = SSE/SST = 1 - SSR/SST$
8. For simple regression it is literally true that $R^2 = r^2$, where r is the simple correlation coefficient between x and y.

OLS Estimation

How do we find the best estimate for b0 and b1?

Method of Moments

1. $E(u) = 0 \Rightarrow E(y - \beta_0 - \beta_1 x) = 0$
2. $E(u|X) = 0 \Rightarrow E[x(y - \beta_0 - \beta_1 x)] = 0$

By (1), $\beta_0 = E(y) - \beta_1 E(x)$, or in terms of the sample $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

By (2), $\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

Substituting for $\hat{\beta}_0$, we get $\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - be\hat{t}a_1 x_i) = 0$

Solving for $\hat{\beta}_1$, we have $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, or $\frac{Cov(xy)}{Var(x)}$.

We can further simplify the first formula for $\hat{\beta}_1$ as $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Ordinary Least Squares:

$$\sum_{i=1}^{n} \hat{u}^2 = \sum_{i=1}^{n} (y_i \breve{\ } \hat{\beta}_0 \breve{\ } \hat{\beta}_1 x_i)^2$$

The first order condition with respect to β_1 is $\sum_{i=1}^{n} 2(y_i \breve{\ } \hat{\beta}_0 \breve{\ } \hat{\beta}_1 x_i)x_i = 0$.

Dividing by 2 and rearranging slightly, $\sum_{i=1}^{n} x_i(y_i \breve{\ } \hat{\beta}_0 \breve{\ } \hat{\beta}_1 x_i) = 0$.

This is identical to the condition for the method of moments estimator above.