Assignment 6b

Satvik Saha

2024-10-14

Answer 1

(a) We might say that comparing two students whose pre-test scores differ by 1, their post-test scores differ by 0.7 on average.

(b) An R^2 score of 1 implies that all residuals are zero. Thus, the data used to fit the model lies precisely on the $y = 30 + 0.7x$ line.

(c) An R^2 score of 0 implies that the sum of squares of residuals is equal to the total sum of squares. This means that the fitted model performs no better than the model which does not look at the pre-test scores at all, and instead predicts the mean of observed post-test scores every time.

Answer 2

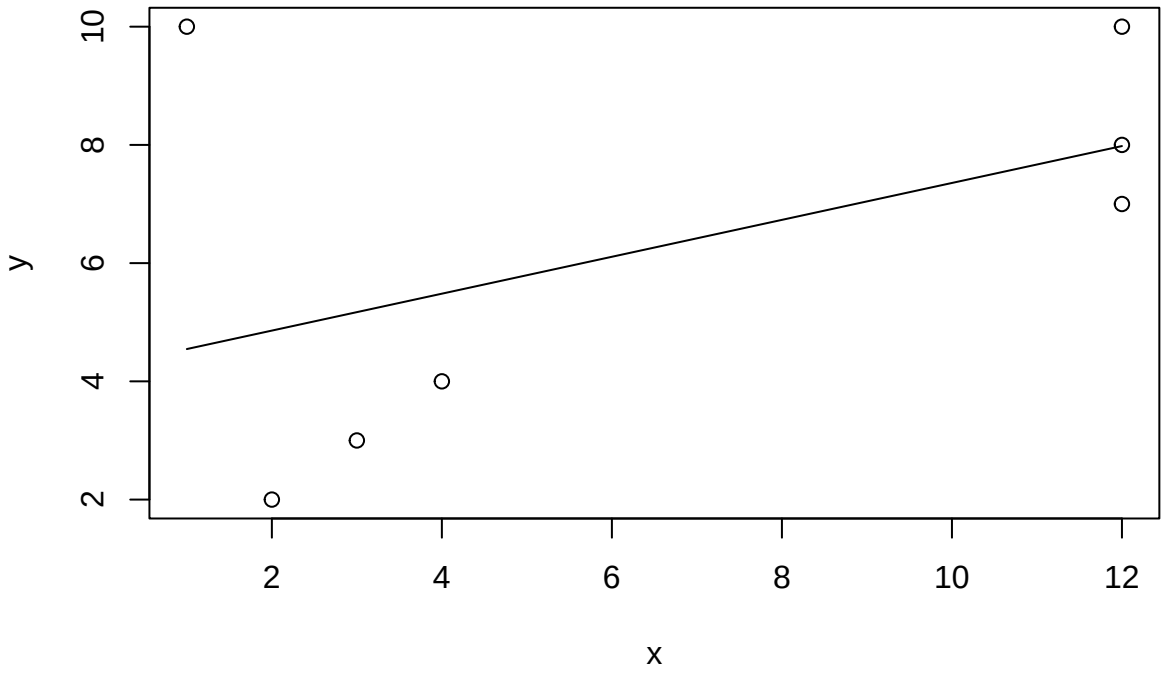
(a) The model can be described as `vote = 46.3 + 3.0 growth + error`. This says that comparing two elections where the average recent growth in personal income differ by 1 percentage point, the incumbent party's vote share differs by 3 percentage points on average.

(b) It would be improper to make the (causal) interpretation that each percentage point of economic growth increases the incumbent party's vote share by 3 percentage points. Our model can only tell us about averages, based on past elections. To assert how much an increase in economic growth will affect an individual election, we must take into account *how* this increase will come about, which could happen in many ways that would actually affect the vote shares differently.

Research homework assignment

```r
x <- c( 1,  2,  3,  4, 12, 12, 12)
y <- c(10,  2,  3,  4,  7,  8, 10)
df <- data.frame(x = x, y = y)

fit <- lm(y ~ x, data = df)
b.hat <- coef(fit)[2]
plot(df)
lines(x, predict(fit, df))
```




```
print(fit)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Coefficients:
## (Intercept)            x
##      4.2343       0.3122
```

```
for (i in 1:nrow(df)) {
    fit.new <- lm(y ~ x, data = df[-i, ])
    df$sens_remove[i] <- coef(fit.new)[2] - b.hat
}
```

Now suppose that the likelihood is tweaked as

$$\mathcal{L}(\beta, \sigma^2; y) = \prod_{i=1}^{n} \left[\phi\left(\frac{y_i - x\beta}{\sigma} \right) \right]^{1+\delta_{ik}\epsilon} = \left[\prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma} \right) \right] \cdot \phi\left(\frac{y_i - x\beta}{\sigma} \right)^{\epsilon}.$$

for some k. The maximum likelihood estimator will minimize

$$\frac{n}{2} \log 2\pi + \frac{n+\epsilon}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (1 + \delta_{ik}\epsilon)(y_i - x\beta)^2.$$

Thus, finding $\hat{\beta}$ reduces to a weighted least squares estimate!

```
epsilon <- 1e-6

for (i in 1:nrow(df)) {
    w <- rep(1, nrow(df))
    w[i] <- 1 + epsilon
    fit.new <- lm(y ~ x, data = df, weights = w)
    df$sens_scale[i] <- (coef(fit.new)[2] - b.hat) / epsilon
}
```

Of course, it is possible to obtain a closed form expression for $d\hat{\beta}_1/d\epsilon$ at $\epsilon = 0$ using the weighted least squares formula. We choose to trust the numerical approximation above since it seems fairly stable.

Finally, suppose that the response y is tweaked as $y + \epsilon e_k$ for some k. Then, $\hat{\beta}_1 = e_2^\top (X^\top X)^{-1} X^\top y$ tells us that $\partial\hat{\beta}_1/\partial y_k$ is just the k-th component of $e_2^\top (X^\top X)^{-1} X^\top$, crucially independent of y. In fact, it is precisely $(x_k - \bar{x})/\sum_i (x_i - \bar{x})^2$, proportional to the deviation $x_k - \bar{x}$. This is just what we have termed the influence of the k-th point.

```
epsilon <- 1e-6

for (i in 1:nrow(df)) {
    y.new <- y
    y.new[i] <- y[i] + epsilon
    df.new <- data.frame(x = x, y = y.new)
    fit.new <- lm(y ~ x, data = df.new)
    df$sens_shift[i] <- (coef(fit.new)[2] - b.hat) / epsilon
}
```

With this, we have our different measures of sensitivity below.

```
df
```

```
##    x  y   sens_remove    sens_scale  sens_shift
## 1  1 10  0.2870257038 -0.1902379022 -0.03488372
## 2  2  2 -0.1126571904  0.0818225525 -0.02862254
## 3  3  3 -0.0624522666  0.0485429406 -0.02236136
## 4  4  4 -0.0292698428  0.0238766475 -0.01610018
## 5 12  7  0.0495375481 -0.0333204150  0.03398927
## 6 12  8 -0.0009943668  0.0006688404  0.03398927
## 7 12 10 -0.1020581966  0.0686473513  0.03398927
```

The fact that the first two measures of sensitivity roughly agree is perhaps not too surprising, given that they both involve some kind of weighting (a weight of 0 for removal). The signs conflict since we are under-weighting in one case, over-weighting in the other. Note that these do depend on the response y; comparing the last three data points (all at $x = 12$), the slope has very low sensitivity with respect to the point with $y = 8$, which is natural given that the regression line almost passes right through it!

The third measure of sensitivity is simply the influence, proportional to the deviation of x from the mean $\bar{x}$ for each data point.