Removal of statistically significant intercept term increases R^2 in linear model

In a simple linear model with a single explanatory variable,

$$\alpha_i = \beta_0 + \beta_1 \delta_i + \epsilon_i$$

I find that removing the intercept term improves the fit greatly (value of R^2 goes from 0.3 to 0.9). However, the intercept term appears to be statistically significant.

With intercept:

```
Call:
lm(formula = alpha ~ delta, data = cf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72138 -0.15619 -0.03744  0.14189  0.70305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.48408    0.05397    8.97   <2e-16 ***
delta        0.46112    0.04595   10.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2435 on 218 degrees of freedom
Multiple R-squared: 0.316,    Adjusted R-squared: 0.3129
F-statistic: 100.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

Without intercept:

```
Call:
lm(formula = alpha ~ 0 + delta, data = cf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92474 -0.15021  0.05114  0.21078  0.85480

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
delta  0.85374    0.01632   52.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2842 on 219 degrees of freedom
Multiple R-squared: 0.9259,   Adjusted R-squared: 0.9256
F-statistic:  2738 on 1 and 219 DF,  p-value: < 2.2e-16
```

How would you interpret these results? Should an intercept term be included in the model or not?

Edit

Here's the residual sums of squares:

```
RSS(with intercept) = 12.92305
RSS(without intercept) = 17.69277
```

r | linear-model | interpretation | r-squared | intercept

edited Dec 27 '15 at 14:07

asked Apr 10 '12 at 11:29

Ernest A
677 1 8 13

9 I recall R^2 to be the ratio of explained to total variance ONLY if the intercept is included. Otherwise it can't be derived and loses its interpretation. – Momo Apr 10 '12 at 11:41

@Momo: Good point. I've calculated the residual sums of squares for each model, which seem to suggest that the model with intercept term is a better fit regardless of what R^2 says. – Ernest A Apr 10 '12 at 12:31

3 Well, the RSS has to go down (or at least not increase) when you include an additional parameter. More importantly, much of the standard inference in linear models does not apply when you suppress the intercept (even if it's not statistically significant). – Macro Apr 10 '12 at 13:11

13 What R does when there is no intercept is that it calculates

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$$

instead (notice, no subtraction of the mean in the denominator terms). This makes the denominator larger which, for the same or similar MSE causes the R^2 to increase. – cardinal ♦ Apr 10 '12 at 14:07

5 The R^2 is not *necessarily* larger. It's only larger without an intercept as long as the MSE of the fit in both cases are similar. But, note that as @Macro pointed out, the numerator *also* gets larger in the case with no intercept so it depends on which one wins out! You're correct that they shouldn't be compared to one another but you *also* know that the SSE with intercept will *always* be smaller than the SSE without intercept. This is part of the problem with using in-sample measures for regression diagnostics. What is your end goal for the use of this model? – cardinal ♦ Apr 10 '12 at 14:36

2 Answers

First of all, we should understand what the `R` software is doing when no intercept is included in the model. Recall that the usual computation of R^2 when an intercept is present is

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \, .$$

The first equality *only* occurs because of the inclusion of the intercept in the model *even though* this is probably the more popular of the two ways of writing it. The *second* equality actually provides the more general interpretation! This point is also address in this related question.

But, what happens if there is no intercept in the model?

Well, in that case, `R` (**silently!**) uses the modified form

$$R_0^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2} \, .$$

It helps to recall what R^2 is trying to measure. In the former case, it is comparing your current model to the *reference* model that only includes an intercept (i.e., constant term). In the second case, there is no intercept, so it makes little sense to compare it to such a model. So, instead, R_0^2 is computed, which implicitly uses a reference model corresponding to **noise only**.

In what follows below, I focus on the second expression for both R^2 and R_0^2 since that expression generalizes to other contexts and it's generally more natural to think about things in terms of residuals.

But, how are they different, and when?

Let's take a brief digression into some linear algebra and see if we can figure out what is going on. First of all, let's call the fitted values from the model *with* intercept $\hat{\mathbf{y}}$ and the fitted values from the model *without* intercept $\tilde{\mathbf{y}}$.

We can rewrite the expressions for R^2 and R_0^2 as

$$R^2 = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2} \, ,$$

and

$$R_0^2 = 1 - \frac{\|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2}{\|\mathbf{y}\|_2^2} \, ,$$

respectively.

Now, since $\|\mathbf{y}\|_2^2 = \|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2 + n\bar{y}^2$, then $R_0^2 > R^2$ if and only if

$$\frac{\|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2} < 1 + \frac{\bar{y}^2}{\frac{1}{n}\|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2} \, .$$

The left-hand side is greater than one since the model corresponding to $\tilde{\mathbf{y}}$ is nested within that of $\hat{\mathbf{y}}$. The second term on the right-hand side is the squared-mean of the responses divided by the mean square error of an intercept-only model. So, the larger the mean of the response relative to the other variation, the more "slack" we have and a greater chance of R_0^2 dominating R^2.

Notice that all the model-dependent stuff is on the left side and non-model dependent stuff is on the right.

Ok, so how do we make the ratio on the left-hand side small?

Recall that $\tilde{\mathbf{y}} = \mathbf{P}_0\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{P}_1\mathbf{y}$ where $\mathbf{P}_0$ and $\mathbf{P}_1$ are projection matrices corresponding to subspaces S_0 and S_1 such that $S_0 \subset S_1$.

So, in order for the ratio to be close to one, we need the subspaces S_0 and S_1 to be very similar. Now S_0 and S_1 differ only by whether $\mathbf{1}$ is a basis vector or not, so that means that S_0 had better be a subspace that already lies very close to $\mathbf{1}$.

In essence, that means our predictor had better have a strong mean offset itself and that this mean offset should dominate the variation of the predictor.

An example

Here we try to generate an example with an intercept explicitly in the model and which behaves close to the case in the question. Below is some simple `R` code to demonstrate.

```
set.seed(.Random.seed[1])

n <- 220
a <- 0.5
b <- 0.5
se <- 0.25

# Make sure x has a strong mean offset
x <- rnorm(n)/3 + a
y <- a + b*x + se*rnorm(x)

int.lm   <- lm(y~x)
noint.lm <- lm(y~x+0)  # Intercept be gone!

# For comparison to summary(.) output
rsq.int <- cor(y,x)^2
rsq.noint <- 1-mean((y-noint.lm$fit)^2) / mean(y^2)
```

This gives the following output. We begin with the model *with* intercept.

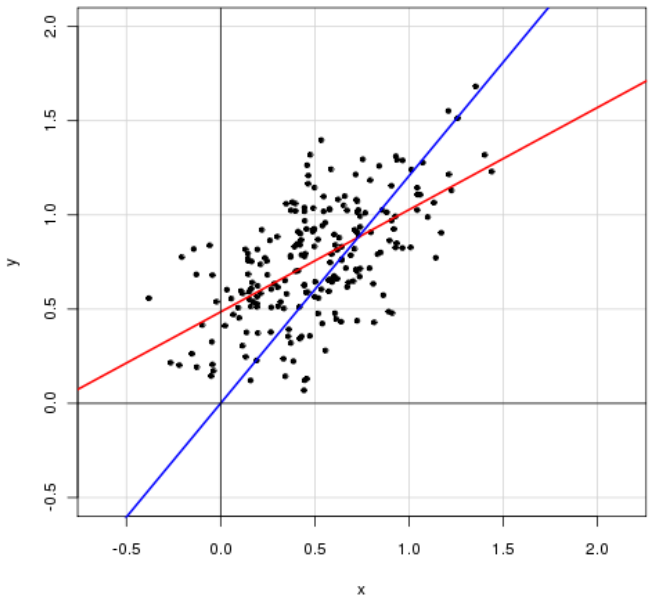
```
# Include an intercept!
> summary(int.lm)

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.656010 -0.161556 -0.005112  0.178008  0.621790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.48521    0.02990   16.23   <2e-16 ***
x            0.54239    0.04929   11.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2467 on 218 degrees of freedom
Multiple R-squared: 0.3571,     Adjusted R-squared: 0.3541
F-statistic: 121.1 on 1 and 218 DF,  p-value: < 2.2e-16
```

Then, see what happens when we *exclude* the intercept.

```
# No intercept!
> summary(noint.lm)

Call:
lm(formula = y ~ x + 0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62108 -0.08006  0.16295  0.38258  1.02485

Coefficients:
```

```
  Estimate Std. Error t value Pr(>|t|)
x  1.20712    0.04066   29.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3658 on 219 degrees of freedom
Multiple R-squared: 0.801,      Adjusted R-squared: 0.8001
F-statistic: 881.5 on 1 and 219 DF,  p-value: < 2.2e-16
```

Below is a plot of the data with the model-with-intercept in red and the model-without-intercept in blue.



edited Apr 13 at 12:44 — Community ♦ — 1

answered Apr 10 '12 at 16:41 — cardinal ♦ — 19.4k 6 73 104

2 This is really spectacular, +1. A question: often when we want to adjudicate b/t 2 models, we perform a nested model test, does this preclude that, or would it still be valid to test a reduced model w/o the intercept against a full model w/ it? – gung ♦ Apr 10 '12 at 19:01

5 @gung: No, I don't believe there is anything precluding us from the typical F-test. The test does not depend on the presence of an intercept, and, indeed, I believe the F-statistic will work out in this case to be

$$F = (n-2)\left(\frac{\|\mathbf{y}-\tilde{\mathbf{y}}\|_2^2}{\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2} - 1\right)$$

. This gives us a bit of quantitative information in that, if indeed $R_0^2 > R_1^2$, then we know that

$$F < (n-2)\frac{\bar{y}^2}{n^{-1}\|\mathbf{y}-\bar{y}\mathbf{1}\|_2^2},$$

assuming, of course, I've done the algebra correctly. – cardinal ♦ Apr 10 '12 at 20:51

1 I prefer the expression $R_0^2 = \frac{\|\tilde{Y}\|^2}{\|Y\|^2}$ – Stéphane Laurent Apr 11 '12 at 4:56

3 @naught101: I would not say it is *more* true, but it is an equally reasonable viewpoint, generally speaking. For the present exposition, it is convenient to consider it as absent in the sense that we are ultimately interested in the relationship between the subspaces S_1 and S_0. The difference between the two is the presence, or lack thereof, of the basis vector $\mathbf{1}$. – cardinal ♦ May 10 '12 at 3:59

1 I'm missing something. Is what R does, *correct*? I mean is the R^2 value that is reported, even remotely comparable between the with and without intercept cases? – Andy Clifton Apr 29 '14 at 0:53

I would base my decision on an information criteria such as the Akaike or Bayes-Schwarz

criteria rather than R^2; even then I would not view these as absolute.

If you have a process where the slope is near zero and all of the data is far from the origin, your correct R^2 should be low as most of the variation in the data will be due to noise. If you try to fit such data to a model without an intercept you will generate a large and wrong slope term and likely a better looking R^2 if the intercept free version is used.

The following graph shows what happens in this extreme cases. Here the generating process is that x=100,100.1,.... and y is just 100 + random noise with mean 0 and standard deviation .1. The points are black circles, the fit without the intercept is the blue line and the fit with the intercept (zeroing out the slope) is the red line:

[Sorry it won't let me post the graph; run the R-code below to generate it. It shows the origin in the lower left corner, the cluster of points in the upper right corner. The bad no-intercept fit goes from the lower left to the upper right and the correct fit is a line parallel to the x-axis]

The correct model for this should have an R^2 of zero---be a constant plus random noise. R will give you and R^2 of .99 for the fit with no intercept. This won't matter much if you only use the model for prediction with x-values within the range of the training data, but will fail miserably if x goes outside of the narrow range of the training set or you are trying to gain true insights beyond just prediction.

The AIC correctly shows that the model with the intercept is preferred. The R code for this is:

```
 Nsamp=100
x=seq(1,100,1)*.1+100 # x=101.1,101.2,....
y=rnorm(n=length(x))+100 # random noise +100 (best model is constant)

model_withint=lm(y~x)
print(summary(model_withint))
flush.console()
model_noint=lm(y~x+0)
print(summary(model_noint))
print (AIC(model_withint))
print(sprintf ('without intercept  AIC=%f',AIC(model_noint)))
print(sprintf ('with intercept  AIC=%f',AIC(model_withint)))
print(sprintf ('constant model  AIC=%f',AIC(lm(y~1))))
plot(x,y,ylim=c(0,105),xlim=c(0,105))
lines( c(0,105),c(0,105)*model_noint$coefficients['x'],col=c('blue'))
lines( c(0,105),c(1,1)*(lm(y~1)$coefficients['(Intercept)']),col=c('red'))
```

The AIC output is

```
   "without intercept  AIC=513.549626"
    "with intercept  AIC=288.112573"
    "constant model  AIC=289.411682"
```

Note that the AIC still gets the wrong model in this case, as the true model is the constant model; but other random numbers will yield data for which the AIC is lowest for the constant model. Note that if you discard the slope, you should refit the model without it, not try to use the intercept from the model and ignore the slope.

edited Aug 17 at 18:00
Nick Cox
32.5k 4 64 103

answered Apr 28 at 16:28
Jonathan Harris
11 2