How Regression is Affected by Correlation of the Predictor Variables

Suppose we do a regression of a response variable, y, on two predictor variables, x_1 and x_2.

How would we expect the results to differ from regressions of y on just x_1 or on just x_2?

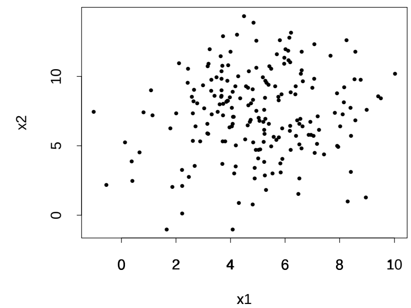
The answer depends a lot on whether x_1 and x_2 are correlated.

High correlation between predictor variables can arise when they are measuring similar things, or are both related to something else. For example:

- Grades in high school and score on an achievement test are both measures of academic achievement.
- Televisions per person and physicians per person may both be related to the wealth of the country.

Example With Uncorrelated Predictors

Here's some artificial data for 200 cases where x_1 and x_2 are nearly uncorrelated, as shown:



I let $y = 8 + 3x_1 - 5x_2 + \varepsilon$, with $\sigma_\varepsilon = 1$.

I then did regressions using MINITAB for y on x_1, for y on x_2, and for y on both x_1 and x_2.

Results of the Regressions

```
The regression equation is
y = - 24.2 + 2.09 x1

Predictor        Coef       StDev          T        P
Constant      -24.187       2.829      -8.55    0.000
x1             2.0946      0.5266       3.98    0.000

S = 15.12       R-Sq = 7.4%      R-Sq(adj) = 6.9%
```

```
The regression equation is
y = 21.1 - 4.74 x2

Predictor        Coef       StDev          T        P
Constant       21.061       1.127      18.69    0.000
x2            -4.7382      0.1417     -33.44    0.000

S = 6.093       R-Sq = 85.0%     R-Sq(adj) = 84.9%
```

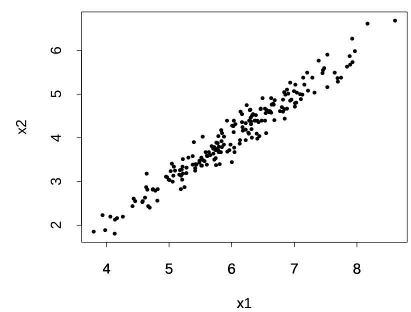
```
The regression equation is
y = 8.01 + 2.97 x1 - 4.97 x2

Predictor        Coef       StDev          T        P
Constant       8.0150      0.2434      32.92    0.000
x1            2.96571     0.03545      83.66    0.000
x2           -4.96981     0.02366    -210.02    0.000

S = 1.011       R-Sq = 99.6%     R-Sq(adj) = 99.6%
```

Example With Correlated Predictors

Now let's look at data for 200 cases where x_1 and x_2 are highly correlated ($r = 0.97$):



This time, I let $y = 4 + 0.7x_1 + \varepsilon$, with $\sigma_\varepsilon = 1$.

Again, I did regressions for y on x_1, for y on x_2, and for y on both x_1 and x_2.

Results of the Regressions

```
The regression equation is
y = 3.76 + 0.746 x1

Predictor        Coef      StDev          T        P
Constant       3.7560     0.4433       8.47    0.000
x1            0.74639    0.07248      10.30    0.000

S = 0.9729      R-Sq = 34.9%     R-Sq(adj) = 34.5%
```

```
The regression equation is
y = 5.31 + 0.730 x2

Predictor        Coef      StDev          T        P
Constant       5.3066     0.3024      17.55    0.000
x2            0.73012    0.07265      10.05    0.000

S = 0.9810      R-Sq = 33.8%     R-Sq(adj) = 33.4%
```

```
The regression equation is
y = 4.03 + 0.603 x1 + 0.146 x2

Predictor        Coef      StDev          T        P
Constant       4.0295     0.7435       5.42    0.000
x1             0.6029     0.3211       1.88    0.062
x2             0.1464     0.3191       0.46    0.647

S = 0.9748      R-Sq = 34.9%     R-Sq(adj) = 34.3%
```

Testing Whether All Regression Coefficients are Zero

When the predictor variables are correlated, it's possible that none of the P-values for tests of whether the coefficients are zero will be significant — even though it's clear that there is a relationship of some sort with one or more of the predictors.

The F test for regression tells us how strong the evidence is that there is a real linear relationship of the response to some predictor or combination of predictors.

As for ANOVA, the F test is derived by analysing how the total sum of squares, SSTotal $= \sum(y_i - \bar{y})^2$, can be partitioned into the part due to error (residuals), SSE, and the part relating to the regression on the predictor variables, SSReg.

ANOVA for Regression

If $\widehat{y}_i = b_0 + b_1 x_{i,1} + \cdots b_k x_{i,k}$ is the value the estimated regression equation predicts for y_i, the sum of squares due to error is

$$\text{SSE} \;=\; \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2$$

The associated "degrees of freedom" is DFE $= n - k - 1$, and the mean square for error is MSE $=$ SSE/DFE, whose square root is s, the estimated residual standard deviation.

The sum of squares due to the regression is

$$\text{SSReg} \;=\; \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \;=\; \text{SSTotal} - \text{SSE}$$

The mean square for regression is MSReg $=$ SSReg/DFReg, with DFReg $= k$.

The proportion of the variability explained by the regression equation is

$$R^2 \;=\; \text{SSReg} \,/\, \text{SSTotal}$$

The F Test for Regression

To test $H_0 : \beta_1 = \cdots = \beta_k = 0$ versus the alternative that at least one coefficient is non-zero, we use the test statistic

$$F \;=\; \text{MSReg} \,/\, \text{MSE}$$

If H_0 is true, this has the F distribution with degrees of freedom DFReg and DFE. We use this distribution to find a P-value from the observed value of F.

Here's the ANOVA and F test for the example:

```
The regression equation is
y = 4.03 + 0.603 x1 + 0.146 x2

Predictor        Coef      StDev          T        P
Constant       4.0295     0.7435       5.42    0.000
x1             0.6029     0.3211       1.88    0.062
x2             0.1464     0.3191       0.46    0.647

S = 0.9748      R-Sq = 34.9%     R-Sq(adj) = 34.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     100.561      50.281     52.91    0.000
Residual Error   197     187.206       0.950
Total            199     287.767
```