13 Multiple Linear Regression

Chapter 12

Multiple Regression Analysis

Definition

The **multiple regression model equation** is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.

Again, it is assumed that ε is normally distributed.

This is not a regression line any longer, but a *regression surface* and we relate *y* to more than one predictor variable $x_1, x_2, \ldots, x_p$. (ex. Blood sugar level vs. weight and age)

Multiple Regression Analysis

The regression coefficient β_1 is interpreted as the expected change in *Y* associated with a 1-unit increase in x_1 *while* x_2,..., x_p *are held fixed*.

Analogous interpretations hold for β_2,..., β_p.

Thus, these coefficients are called *partial* or *adjusted regression coefficients*.

In contrast, the *simple regression* slope is called the marginal (or *unadjusted*) coefficient.

Easier Notation?

The multiple regression model can be written in matrix form.

Estimating Parameters

To estimate the parameters β_0, β_1,..., β_p using the principle of least squares, form the sum of squared deviations of the observed y_j's from the regression line:

$$\mathcal{Q} = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (Y_i - (\beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{pi}))^2$$

The least squares estimates are those values of the β_js that minimize the equation. You *could* do this by taking the partial derivative w.r.t. to each parameter, and then solving the *k+1* unknowns using the *k+1* equations (akin to the simple regression method).

But we don't do it that way.

Models with Categorical Predictors

Sometimes, a three-category variable can be included in a model as one covariate, coded with values 0, 1, and 2 (or something similar) corresponding to the three categories.

This is generally incorrect, because it imposes an ordering on the categories that may not exist in reality. Sometimes it's ok to do this for education categories (e.g., HS=1,BS=2,Grad=3), but not for ethnicity, for example.

The correct approach to incorporating three unordered categories is to define *two* different indicator variables.

Example

Suppose, for example, that y is the lifetime of a certain tool, and that there are 3 brands of tool being investigated.

Let:

$x_1 = 1$ if tool A is used, and 0 otherwise,

$x_2 = 1$ if tool B is used, and 0 otherwise,

$x_3 = 1$ if tool C is used, and 0 otherwise.

Then, if an observation is on a:

brand A tool: we have $x_1 = 1$ and $x_2 = 0$ and $x_3 = 0$,

brand B tool: we have $x_1 = 0$ and $x_2 = 1$ and $x_3 = 0$,

brand C tool: we have $x_1 = 0$ and $x_2 = 0$ and $x_3 = 1$.

What would our X matrix look like?

R^2 and $\hat{\sigma}^2$

R^2

Just as with simple regression, the error sum of squares is

$$SSE = \Sigma(y_i - \hat{y}_i)^2.$$

It is again interpreted as a measure of how much variation in the observed *y* values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is $n-(p+1)$ because $p+1$ df are lost in estimating the $p+1$ β coefficients.

R^2

Just as before, the total sum of squares is

$$\text{SST} = \Sigma(y_i - \bar{y})^2,$$

And the regression sum of squares is:

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2 = SST - SSE.$$

Then the **coefficient of multiple determination R^2** is

$$R^2 = 1 - \text{SSE/SST} = \text{SSR/SST}$$

It is interpreted in the same way as before.

R^2

Unfortunately, there is a problem with R^2: Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous.

R^2

For example, suppose y is the sale price of a house. Then sensible predictors include

x_1 = the interior size of the house,

x_2 = the size of the lot on which the house sits,

x_3 = the number of bedrooms,

x_4 = the number of bathrooms, and

x_5 = the house's age.

Now suppose we add in

x_6 = the diameter of the doorknob on the coat closet,

x_7 = the thickness of the cutting board in the kitchen,

x_8 = the thickness of the patio slab.

R^2

The objective in multiple regression is not simply to explain most of the observed *y* variation, but to do so using a model with relatively few predictors that are easily interpreted.

It is thus desirable to adjust R^2 to take account of the size of the model:

$$R_a^2 = 1 - \frac{SSE/\big(n-(p+1)\big)}{SST/(n-1)} = 1 - \frac{n-1}{n-(p+1)} \times \frac{SSE}{SST}$$

R^2

Because the ratio in front of SSE/SST exceeds 1, R_a^2 is smaller than R^2. Furthermore, the larger the number of predictors p relative to the sample size n, the smaller R_a^2 will be relative to R^2.

Adjusted R^2 can even be negative, whereas R^2 itself must be between 0 and 1. A value of R_a^2 that is substantially smaller than R^2 itself is a warning that the model may contain too many predictors.

$\hat{\sigma}^2$

SSE is still the basis for estimating the remaining model parameter:

$$\widehat{\sigma^2} = s^2 = \frac{SSE}{n-(p+1)}$$

Example

Investigators carried out a study to see how various characteristics of concrete are influenced by

x_1 = % limestone powder

x_2 = water-cement ratio,

resulting in data published in "Durability of Concrete with Addition of Limestone Powder," *Magazine of Concrete Research*, 1996: 131–137.

Example

cont'd

Consider predicting compressive strength (`strength`) with percent limestone powder (`perclime`) and water-cement ratio (`watercement`).

```
> fit = lm(strength ~ perclime + watercement, data = dataset)
> summary(fit)
...
Coefficients:            Estimate Std. Error t value Pr(>|t|)
(Intercept)               86.2471    21.7242   3.970  0.00737 **
perclime                   0.1643     0.1993   0.824  0.44119
watercement              -80.5588    35.1557  -2.291  0.06182 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.832 on 6 degrees of freedom
Multiple R-squared:  0.4971,  Adjusted R-squared:  0.3295
F-statistic: 2.965 on 2 and 6 DF,  p-value: 0.1272
```

Example

Now what happens if we add an interaction term? How do we interpret this model?

```
> fit.int = lm(strength ~ perclime + watercement +
perclime:watercement, data = dataset)
> summary(fit.int)
...
Coefficients:              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.647     56.492   0.135    0.898
perclime                    5.779      3.783   1.528    0.187
watercement                50.441     93.821   0.538    0.614
perclime:watercement       -9.357      6.298  -1.486    0.197
Residual standard error: 4.408 on 5 degrees of freedom
Multiple R-squared:  0.6511,  Adjusted R-squared:  0.4418
F-statistic: 3.111 on 3 and 5 DF,  p-value: 0.1267
```

Model Selection

Important Questions:

- Model utility: Are all predictors significantly related to our outcome? (Is our model any good?)
- Does any particular predictor or predictor subset matter more?
- Are any predictors related to each other?
- Among all possible models, which is the “best”?

A Model Utility Test

The model utility test in simple linear regression involves the null hypothesis H_0: $\beta_1 = 0$, according to which there is no useful linear relation between *y* and the predictor *x.*

In MLR we test the hypothesis

$$H_0: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0,$$

which says that there is no useful linear relationship between *y* and *any* of the *p* predictors. If at least one of these β's is not 0, the model is deemed useful.

We *could* test each β separately, but that would take time and be very conservative (if Bonferroni correction is used). **A better test is a joint test, and is based on a statistic that has an *F* distribution when H_0 is true**.

A Model Utility Test

Null hypothesis: H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

Alternative hypothesis: H_a: at least one $\beta_i \neq 0$ ($i = 1,\ldots, p$)

Test statistic value:

$$f = \frac{SSR/p}{SSE/(n-(p+1))}$$

Rejection region for a level α test: $f \geq F_{\alpha,p,n-(p+1)}$

Example – Bond shear strength

The article “How to Optimize and Control the Wire Bonding Process: Part II” (*Solid State Technology*, Jan 1991: 67-72) described an experiment carried out to asses the impact of force (gm), power (mW), temperature (C) and time (msec) on ball bond shear strength (gm).

Example – Bond shear strength

The article “How to Optimize and Control the Wire Bonding Process: Part II” (*Solid State Technology*, Jan 1991: 67-72) described an experiment carried out to asses the impact of force (gm), power (mW), temperature (C) and time (msec) on ball bond shear strength (gm). The output for this model looks like this:

```
Coefficients:      Estimate Std. Error t value Pr(>|t|)
(Intercept) -37.42167   13.10804  -2.855  0.00853 **
force         0.21083    0.21071   1.001  0.32661
power         0.49861    0.07024   7.099 1.93e-07 ***
temp          0.12950    0.04214   3.073  0.00506 **
time          0.25750    0.21071   1.222  0.23308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.161 on 25 degrees of freedom
Multiple R-squared:  0.7137,  Adjusted R-squared:  0.6679
F-statistic: 15.58 on 4 and 25 DF,  p-value: 1.607e-06
```

Example – Bond shear strength

How do we interpret our model results?

Example – Bond shear strength

A model with $p = 4$ predictors was fit, so the relevant hypothesis to determine if our model is "okay" is

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a: at least one of these four βs is not 0

In our output, we see:

```
Coefficients:      Estimate Std. Error t value Pr(>|t|)
(Intercept) -37.42167   13.10804  -2.855  0.00853 **
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.161 on 25 degrees of freedom
Multiple R-squared:  0.7137,     Adjusted R-squared:  0.6679
F-statistic: 15.58 on 4 and 25 DF,  p-value: 1.607e-06
```

Example – Bond shear strength

The null hypothesis should be rejected at any reasonable significance level.

We conclude that there *is* a useful linear relationship between *y* and ***at least one*** of the four predictors in the model.

This does not mean that all four predictors are useful!

Inference for Single Parameters

All standard statistical software packages compute and show the standard deviations of the regression coefficients.

Inference concerning a single β_i is based on the standardized variable

$$T = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}}$$

which has a *t* distribution with $n - (p + 1)$ df. A $100(1 - \alpha)\%$ CI for β_i is

$$\hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} \cdot s_{\hat{\beta}_i}$$

This is the same thing we did for simple linear regression.

Inference for Single Parameters

Our output:

```
Coefficients:      Estimate Std. Error t value Pr(>|t|)
(Intercept) -37.42167   13.10804  -2.855  0.00853 **
force         0.21083    0.21071   1.001  0.32661
power         0.49861    0.07024   7.099 1.93e-07 ***
temp          0.12950    0.04214   3.073  0.00506 **
time          0.25750    0.21071   1.222  0.23308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
...
```

What is the difference between testing each of these parameters individually and our F-test from before?

Inference for Parameter Subsets

In our output, we see that perhaps “force” and “time” can be deleted from the model. We then have these results:

```
Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)   -24.89250   10.07471  -2.471  0.02008 *
power           0.49861    0.07088   7.035 1.46e-07 ***
temp            0.12950    0.04253   3.045  0.00514 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.208 on 27 degrees of freedom
Multiple R-squared:  0.6852,  Adjusted R-squared:  0.6619
F-statistic: 29.38 on 2 and 27 DF,  p-value: 1.674e-07
```

Inference for Parameter Subsets

In our output, we see that perhaps “force” and “time” can be deleted from the model. We then have these results:

```
Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)   -24.89250   10.07471  -2.471  0.02008 *
power           0.49861    0.07088   7.035 1.46e-07 ***
temp            0.12950    0.04253   3.045  0.00514 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.208 on 27 degrees of freedom
Multiple R-squared:  0.6852,  Adjusted R-squared:  0.6619
F-statistic: 29.38 on 2 and 27 DF,  p-value: 1.674e-07
```

In our previous model:

Multiple R-squared: 0.7137, Adjusted R-squared: 0.6679

Inference for Parameter Subsets

An *F* Test for a Group of Predictors.

The "model utility *F* test" was appropriate for testing whether there is useful information about the dependent variable in *any* of the *p* predictors (i.e., whether $\beta_1 = \ldots = \beta_p = 0$).

In many situations, one first builds a model containing *p* predictors and then wishes to know whether any of the predictors in a *particular subset* provide useful information about *Y*.

Inference for Parameter Subsets

The relevant hypothesis is then:

H_0: $\beta_{l+1} = \beta_{l+2} = \cdots = \beta_{l+k} = 0$,

H_a: at least one among $\beta_{l+1}, \ldots, \beta_{l+k}$ is not 0.

Inferences for Parameter Subsets

The test is carried out by fitting both the full and reduced models.

Because the full model contains not only the predictors of the reduced model but also some extra predictors, it should fit the data at least as well as the reduced model.

That is, if we let SSE_p be the sum of squared residuals for the full model and SSE_k be the corresponding sum for the reduced model, then $SSE_p \leq SSE_k$.

Inferences for Parameter Subsets

Intuitively, if SSE_p is a great deal smaller than SSE_k, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction $SSE_k - SSE_p$ in unexplained variation.

SSE_p = unexplained variation for the full model

SSE_k = unexplained variation for the reduced model

Test statistic value: $f = \dfrac{(SSE_k - SSE_p)/(p-k)}{SSE_p/(n-(p+1))}$

Rejection region: $f \geq F_{\alpha,p-k,n-(p+1)}$

Inferences for Parameter Subsets

Let's do this for the bond strength example:

```
> anova(fitfull)
Analysis of Variance Table
Response: strength           Df  Sum Sq Mean Sq F value    Pr(>F)
force       1   26.67   26.67  1.0012  0.326611
power       1 1342.51 1342.51 50.3967 1.931e-07 ***
temp        1  251.55  251.55  9.4431  0.005064 **
time        1   39.78   39.78  1.4934  0.233080
Residuals 25  665.97   26.64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(fitred)
Analysis of Variance Table
Response: strength           Df  Sum Sq Mean Sq F value    Pr(>F)
power       1 1342.51 1342.51 49.4901 1.458e-07 ***
temp        1  251.55  251.55  9.2732  0.005142 **
Residuals 27  732.43   27.13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inferences for Parameter Subsets

Let's do this for the bond strength example:

```
> anova(fitfull, fitred)
Analysis of Variance Table
Model 1: strength ~ force + power + temp + time
Model 2: strength ~ power + temp
      Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     25 665.97
2     27 732.43 -2      -66.454     1.2473 0.3045
```

Multicollinearity

What is multicollinearity?

Multicollinearity occurs when 2 or more predictors in one regression model are highly correlated. Typically, this means that one predictor is a function of the other.

We almost always have multicollinearity in the data. The question is whether we can get away with it; and what to do if multicollinearity is so serious that we cannot ignore it.

Multicollinearity

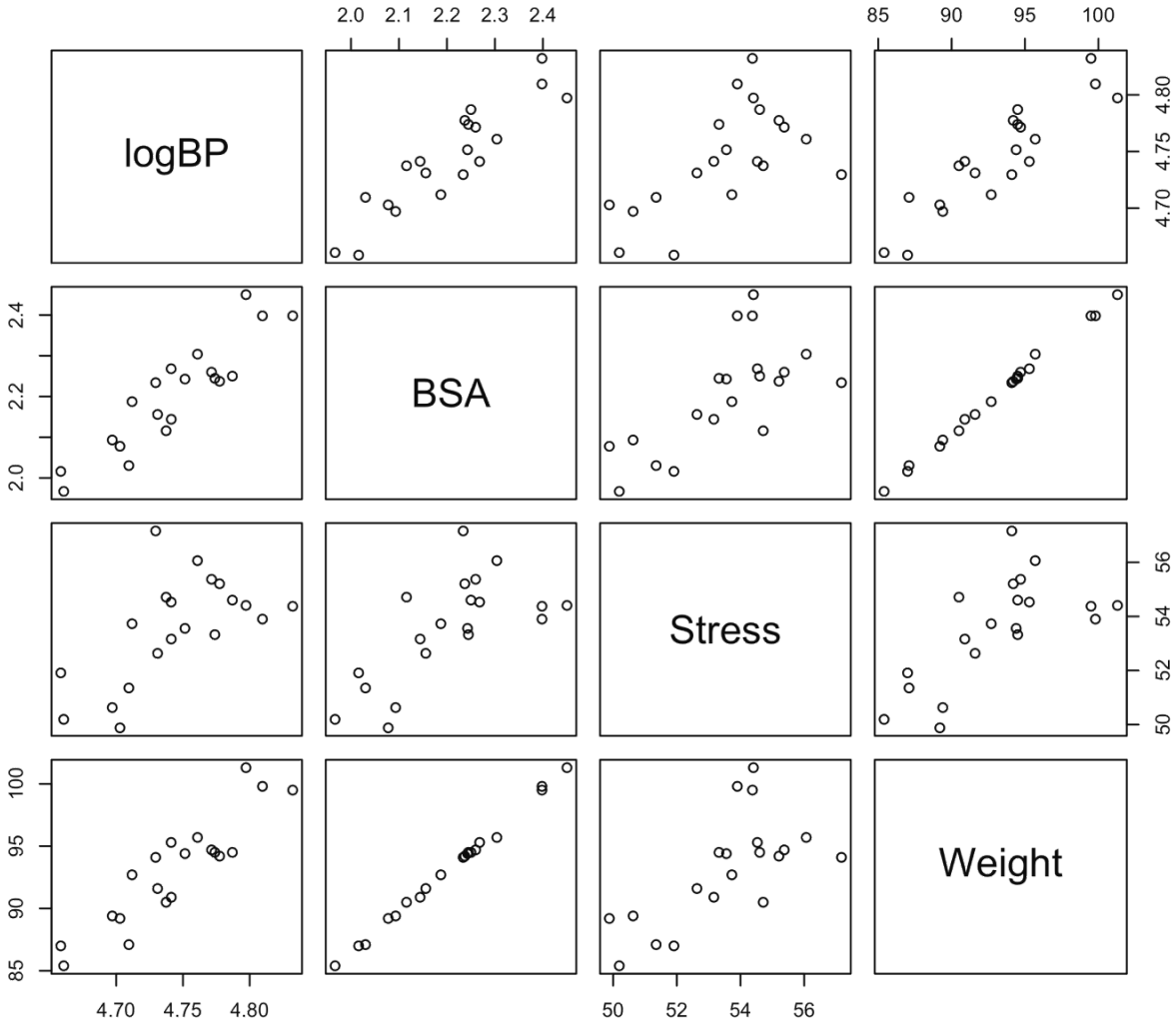
Example: Clinicians observed the following measurements for 20 subjects:

- Blood pressure (in mm Hg)
- Weight (in kg)
- Body surface area (in sq m)
- Stress index

The researchers were interested in determining if a relationship exists between blood pressure and the other covariates.

Multicollinearity

A scatterplot of the predictors looks like this:



Multicollinearity

And the correlation matrix looks like this:

	logBP	BSA	Stress	Weight
logBP	1.000	0.908	0.616	0.905
BSA	0.908	1.000	0.680	0.999
Stress	0.616	0.680	1.000	0.667
Weight	0.905	0.999	0.667	1.000

Multicollinearity

A model summary (including all the predictors, with blood pressure log-transformed) looks like this:

```
Coefficients:   Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.2131301  0.5098890   8.263 3.64e-07 ***
BSA            0.5846935  0.7372754   0.793    0.439
Stress        -0.0004459  0.0035501  -0.126    0.902
Weight        -0.0078813  0.0220714  -0.357    0.726
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.02105 on 16 degrees of freedom
Multiple R-squared:  0.8256,  Adjusted R-squared:  0.7929
F-statistic: 25.25 on 3 and 16 DF,  p-value: 2.624e-06
```

What is Multicollinearity?

The overall F-test has a p-value of `2.624e-06`, indicating that we should reject the null hypothesis that none of the variables in the model are significant.

But none of the individual variables is significant. All p-values are bigger than 0.43.

Multicollinearity may be a culprit here.

Multicollinearity

Multicollinearity is not an error – it comes from the lack of information in the dataset.

For example if

$$X_1 \cong a + b*X_2 + c* X_3,$$

then the data doesn't contain much information about how X_1 varies that isn't already contained in the information about how X_2 and X_3 vary.

Thus we can't have much information about how changing X_1 affects Y if we insist on not holding X_2 and X_3 constant.

Multicollinearity

What happens if we ignore multicollinearity problem?

If it is not "serious", the only thing that happens is that our confidence intervals are a bit bigger than what they would be if all the variables are independent (i.e. all our tests will be slightly more conservative, in favor of the null).

But if multicollinearity is serious and we ignore it, all confidence intervals will be a lot bigger than what they would be, the numerical estimation will be problematic, and the estimated parameters will be all over the place.

This is how we get in this situation when the overall *F-test* is significant, but none of the individual coefficients are.

Multicollinearity

When is multicollinearity serious and how do we detect this?

- Plots and correlation tables show highly linear relationships between predictors.
- A significant F-statistic for the overall test of the model but no single (or very few single) predictors are significant
- The estimated effect of a covariate may have an opposite sign from what you (and everyone else) would expect.

Reducing multicollinearity

STRATEGY 1: Omit redundant variables. (Drawbacks? Information needed?)

STRATEGY 2: Center predictors at or near their mean before constructing powers (square, etc) and interaction terms involving them.

STRATEGY 3: Study the principal components of the *X* matrix to discern possible structural effects (outside of scope of this course).

STRATEGY 4: Get more data with *X's* that lie in the areas about which the current data are not informative (when possible).

Model Selection Methods

So far, we have discussed a number of methods for finding the "best" model:

- Comparison of R^2 and adjusted R^2.
- F-test for model utility and F-test for determining significance of a subset of predictors.
- Individual parameter t-tests.
- Reduction of collinearity.
- Transformations.
- Using your brain.

Model Selection Methods

So far, we have discussed a number of methods for finding the "best" model:

- Comparison of R^2 and adjusted R^2.
- F-test for model utility and F-test for determining significance of a subset of predictors.
- Individual parameter t-tests.
- Reduction of collinearity.
- Transformations.
- **Using your brain.**

Model Selection Methods

So far, we have discussed a number of methods for finding the "best" model:

- Comparison of R^2 and adjusted R^2.
- F-test for model utility and F-test for determining significance of a subset of predictors.
- Individual parameter t-tests.
- Reduction of collinearity.
- Transformations.
- Using your brain.
- Forward and backward stepwise regression, AIC values, etc. (graduate students)