Multiple Regression

Readings: DeVeaux *et al* Chapters 30, 31

1. data consist of measurements on n subjects. For each subject, there is a measurement on the dependent variable y, and on each of q independent variables $x_1, x_2, \ldots, x_q$.
2. The statistical model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \epsilon$$

   Letting $\mu_{y|x}$ stand for "the mean of y, given x", another way to write the model is

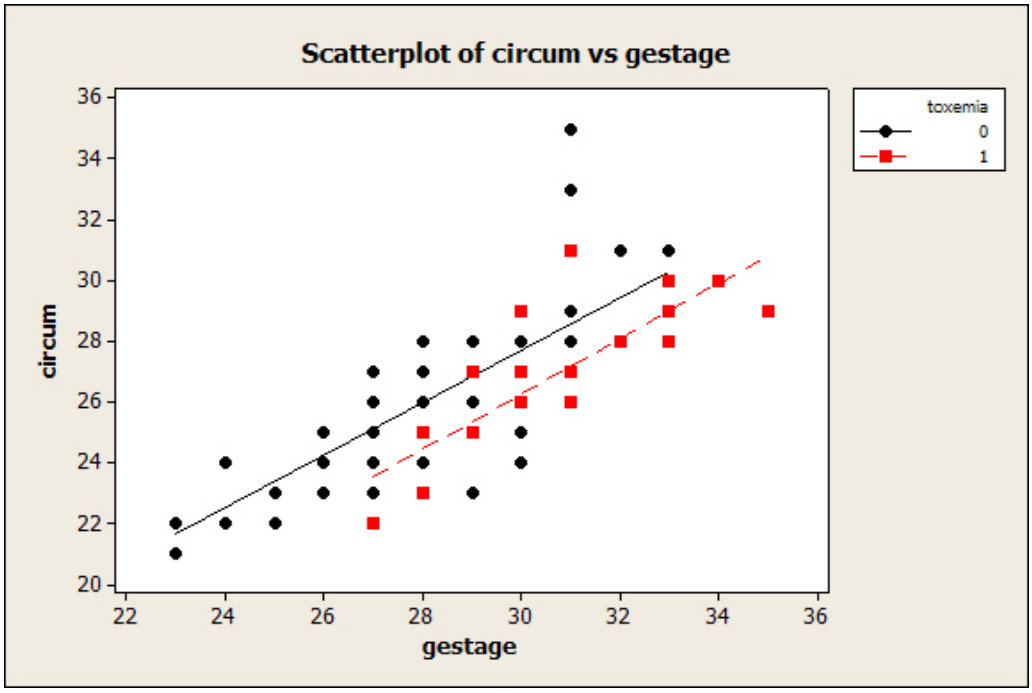
$$y = \mu_{y|x} + \epsilon$$

3. The errors are assumed to be a sample from $N(0, \sigma^2)$.
4. Values for α and $\beta_1, \ldots, \beta_q$ are estimated from the data by the method of least squares.
5. The regression equation, or prediction line, is

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_q x_q$$

circum	length	bwt	gage	mage	toxem	gage×toxem
y	x_1	x_2	x_3	x_4	x_5	$x_6 = x_3 \times x_5$
27	41	1360	29	37	0	0
29	40	1490	31	34	0	0
30	38	1490	33	32	0	0
28	38	1180	31	37	0	0
29	38	1200	30	29	1	30
23	32	680	25	19	0	0
22	33	620	27	20	1	27

Example: We are interested in predicting head circumference on the basis of gestational age, toxemia, and the product of those two variables. The scatter plot below shows data points of 100 infants but due to both gestational age and circumference being rounded to the nearest intergers, many data points are overlapped.



The regression equation is

$$y = \alpha + \beta_1 x_3 + \beta_2 x_5 + \beta_3 (x_3 x_5) + \epsilon$$

This model allows for different slopes and intercepts depending on presence or absence of toxemia.

The following table gives the mean of the regression equation for specified values of parameters.

	no toxemia ($x_5 = 0$)	toxemia ($x_5 = 1$)
	$\alpha + \beta_1 x_3$	$(\alpha + \beta_2) + (\beta_1 + \beta_3) x_3$
$\beta_3 = 0$	$\alpha + \beta_1 x_3$	$(\alpha + \beta_2) + \beta_1 x_3$
$\beta_2 = 0$	$\alpha + \beta_1 x_3$	$\alpha + (\beta_1 + \beta_3) x_3$
$\beta_2 = \beta_3 = 0$	$\alpha + \beta_1 x_3$	$\alpha + \beta_1 x_3$

hypothesis	interpretation
no restrictions	different slopes and intercepts for toxemic vs non-toxemic
$\beta_3 = 0$	same slope, different interecpts for toxemic vs non-toxemic
$\beta_2 = 0$	different slopes, same intercepts for toxemic vs non-toxemic
$\beta_2 = \beta_3 = 0$	no difference in slope or intercept for toxemic vs non-toxemic

A computer program gave the following partial output:

```
Regression Analysis of circumference on gestage, toxemia,
gage*tox

The regression equation is
circum = 1.76 + 0.865 gestage - 2.82 toxemia + 0.046 gage*tox

Predictor      Coef  SE Coef      T      P
Constant      1.763    2.102   0.84  0.404
gestage     0.86461  0.07390  11.70  0.000
toxemia      -2.815    4.985  -0.56  0.574
gage*tox     0.0462   0.1635   0.28  0.778


S = 1.51460   R-Sq = 65.3\%   R-Sq(adj) = 64.2\%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        3  414.53  138.18  60.23  0.000
Residual Error   96  220.22    2.29
Total            99  634.75
```

The computer output gives the least squares estimates together with estimated standard errors. These are used to make confidence intervals and test hypotheses. For example $\hat{\beta}_2 = -2.815$, and the estimated standard error of $\hat{\beta}_2$ is 4.985.

1. **hypothesis tests for individual coefficients**

- To test the hypothesis $H_0 : \beta_j = 0$ against the two sided alternative, calculate the observed value of the test statistic

$$t = \frac{\hat{\beta}_j}{\widehat{s.e.}(\hat{\beta}_j)}$$

- The p-value is $2P(t_{n-1-q} > |t_{obs}|)$.
- eg. for testing $\beta_2 = 0$, the observed test statistic is $t_{obs} = -2.815/4.985 = -.56$. The p-value is $2P(t_{96} > |-.56|) = .574$
- Note that the degrees of freedom of this hypothesis test (and the confidence interval below) is the degrees of freedom for Residual Error.
- When testing the coefficient associated with the variable x_j, the actual hypothesis that is being tested is:

  H_0 : variable x_j has no effect, **given that all all other predictor variables are in the model**

  H_A : variable x_j has an effect, **given that all all other predictor variables are in the model**

  alternatively

  H_0 : variable x_j provides no significant additional reduction in SSE **given that all all other predictor variables are in the model**

  H_A : variable x_j provides a significant additional reduction in SSE, **given that all all other predictor variables are in the model**

2. **Confidence Intervals for individual coefficients**

- A $100(1-\alpha)\%$ CI for β_j is $\hat{\beta}_j \pm t_{\alpha/2,n-1-q}\widehat{s.e.}(\hat{\beta}_j)$
- eg. a 95% CI for β_2 is $\hat{\beta}_2 \pm t_{.025,96}\widehat{s.e.}(\hat{\beta}_2)$, or $-2.815 \pm 1.985(4.985)$, or (-12.71,7.08).

3. **Overall utility of the regression model**

- The F statistic is testing the hypothesis of no effect of any of the variables

  $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$

  H_A: one or more of β_1, β_2, ..., β_q are non-zero
- the p-value for this test is $P(F_{q,n-1-q} > F_{obs})$
- in the example $F_{obs} = 60.23$, and the p-value is $P(F_{3,96} > 60.23) <$ .001 from tables

- The goodness of fit of the linear regression line is measured by the **coefficient of determination**

$$R^2 = \frac{SSR}{SST}$$

  , the ratio of the regresson sum of squares to the total sum of squares.
- R^2 is the fraction of the total variability in y accounted for by the regression line, and ranges between 0 and 1. $R^2 = 1.00$ indicates a perfect (linear) fit, while $R^2 = 0.00$ is a complete lack of linear fit.
- In the example, the linear effects of the three variables gestage, toxemia and gestage $\times$ toxemia (known as the interaction between gestational age and toxemia), account for 65.3% of the variation in head circumference

  An estimate of σ is $\hat{\sigma} = \sqrt{MSE}$, where $MSE = SSE/(n-1-q)$. In the example this is $\sqrt{2.29} = 1.51$.

4. **Model selection**

- One strategy for model selection is to fit a collection of models, and pick the model which has the largest **adjusted** R^2. The adjustment takes into account the number of predictors in the model, and guards against overfitting.

  There are several other criteria that can be used, including Mallow's C_p and Akaike's Information Criterion (AIC).

  Don't choose a model based on R^2, as it always increases when more predictors are included.

- Another strategy is to fit a large model to start, and then remove variables one at a time as the hypothesis tests for single variables dictate. For example, in the above multiple regression, the interaction term is not significant, and in fact, the interaction is the least significant predictor. Therefore, we can remove it from the model.

This leads to a second model

$$y = \alpha + \beta_1 x_3 + \beta_2 x_5 + \epsilon$$

This is a parallel line regression model. There is a common slope, but different intercepts depending on presence or absence of toxemia. In particular, if toxemia=0 ($x_5 = 0$), the model is

$$y = \alpha + \beta_1 x_3 + \epsilon$$

If toxemia=1 ($x_5 = 1$), the model is

$$y = (\alpha + \beta_2) + \beta_1 x_3 + \epsilon$$

```
Regression Analysis of circumference on gestage and toxemia


The regression equation is
circum = 1.50 + 0.874 gestage - 1.41 toxemia

Predictor     Coef  SE Coef      T      P
Constant     1.496    1.868   0.80  0.425
gestage    0.87404  0.06561  13.32  0.000
toxemia    -1.4123   0.4062  -3.48  0.001

S = 1.50739   R-Sq = 65.3\%   R-Sq(adj) = 64.6\%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       2  414.34  207.17  91.18  0.000
Residual Error  97  220.41    2.27
Total           99  634.75
```

- All of the terms in this model (gestage and toxemia) are significant (small associated p-values), so we can stop taking predictive variables out of the model.
- This simpler model still explains 65.3% of the variability in circumference.
- Note that, compared to the model which includes the interaction, R^2 has **no change**, **but** adjusted R^2 has **in**creased.

Here's another example, starting with all predictors, but no interaction variables, and eliminating variables one at a time.

1. The regression equation is

```
circum = 7.21 + 0.0083 length + 0.526 gestage +
0.00426 birthwt - 0.0301 momage  - 0.516 toxemia

Predictor       Coef    SE Coef      T      P
Constant       7.210      2.129   3.39  0.001
length       0.00827    0.06534   0.13  0.900
gestage      0.52619    0.08356   6.30  0.000
birthwt    0.0042555  0.0008867   4.80  0.000
momage      -0.03007    0.02223  -1.35  0.179
toxemia      -0.5161     0.3696  -1.40  0.166

S = 1.26902   R-Sq = 76.2\%   R-Sq(adj) = 74.9\%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       5  483.372  96.674  60.03  0.000
Residual Error  94  151.378   1.610
Total           99  634.750
```

   Baby's length is least important predictor, given that other variables are in the model, and non-significant, so remove at next step.

2. Regression Analysis: circum versus gestage, birthwt, momage, toxemia

   The regression equation is circum = 7.35 + 0.529 gestage + 0.00433 birthwt - 0.0298 momage - 0.516 toxemia

Predictor	Coef	SE Coef	T	P
Constant	7.352	1.800	4.08	0.000
gestage	0.52881	0.08053	6.57	0.000
birthwt	0.0043275	0.0006764	6.40	0.000
momage	-0.02979	0.02201	-1.35	0.179
toxemia	-0.5159	0.3677	-1.40	0.164

S = 1.26243 R-Sq = 76.1\% R-Sq(adj) = 75.1\%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	483.35	120.84	75.82	0.000
Residual Error	95	151.40	1.59		
Total	99	634.75			

Note that R^2 decreased, but adjusted R^2 increased.

Mom's age is now least important, and non-significant, so remove.

3.

```
The regression equation is
circum = 7.10 + 0.508 gestage + 0.00435 birthwt - 0.513 toxemi

Predictor       Coef    SE Coef      T      P
Constant       7.096      1.798   3.95  0.000
gestage      0.50805    0.07940   6.40  0.000
birthwt    0.0043541  0.0006791   6.41  0.000
toxemia      -0.5128     0.3693  -1.39  0.168

S = 1.26789   R-Sq = 75.7\%   R-Sq(adj) = 74.9\%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        3  480.43  160.14  99.62  0.000
Residual Error   96  154.32    1.61
Total            99  634.75
```

Both adjusted R^2 and R^2 have decreased. toxemia is non-significant, so remove.

4. The regression equation is

```
circum = 8.31 + 0.449 gestage + 0.00471 birthwt

Predictor       Coef    SE Coef     T      P
Constant       8.308      1.579  5.26  0.000
gestage      0.44873    0.06725  6.67  0.000
birthwt    0.0047123  0.0006312  7.47  0.000


S = 1.27394   R-Sq = 75.2\%   R-Sq(adj) = 74.7\%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        2  477.33  238.66  147.06  0.000
Residual Error   97  157.42    1.62
Total            99  634.75
```

- All variables are now highly significant.
- Once a model has been selected, good statistical practice dictates the assessment of model assumptions. In this case, that would include assessing the assumption of normality.
- The predicted circumference of a baby with gestational age 30 weeks and a birthweight of 1200 g is

$$8.308 + .44873(30) + .0047123(1200) = 27.43.$$