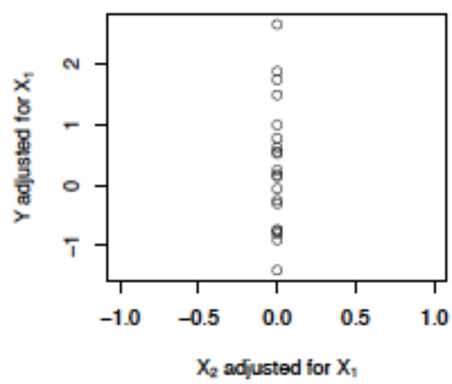
3.3 The following questions all refer to the mean function

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3.26}$$

3.3.1. Suppose we fit (3.26) to data for which $x_1 = 2.2x_2$, with no error. For example, x_1 could be a weight in pounds, and x_2 the weight of the same object in kg. Describe the appearance of the added-variable plot for X_2 after X_1.

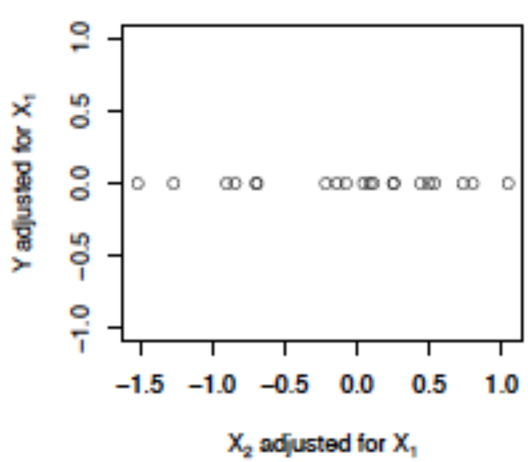
Solution: Since X_2 is an exact linear function of X_1, the residuals from the regression of X_2 on X_1 will all be zero, and so the plot will look like



In general, if X_1 and X_2 are highly correlated, the variability on the horizontal axis of an added-variable plot will be very small compared to the variability of the original variable. The coefficient for such a variable will be very poorly estimated. ∎

3.3.2. Again referring to (3.26), suppose now that Y and X_1 are perfectly correlated, so $Y = 3X_1$, without any error. Describe the appearance of the added-variable plot for X_2 after X_1.

Solution: Since $Y = 3X_1$ the residuals from the regression of Y on X_1 will all be zero, and so the plot will look like



In general, if Y and X_1 are highly correlated, the variability on the vertical axis of an added-variable plot will be very small compared to the variability of the original variable, and we will get an approximately null plot. ∎

3.3.3. Under what conditions will the added-variable plot for X_2 after X_1 have exactly the same shape as the scatterplot of Y versus X_2?

Solution: If X_1 is uncorrelated with both X_2 *and* Y, then these two plots will be the same. ∎

3.3.4. True or false: The vertical variation in an added-variable plot for X_2 after X_1 is always less than or equal to the vertical variation in a plot of Y versus X_2. Explain.

Solution: Since the vertical variable is the residuals from the regression of Y on X_1, the vertical variation in the added-variable plot is never larger than the vertical variation in the plot of Y versus X_2. ∎

3.4 Suppose we have a regression in which we want to fit the mean function (3.1). Following the outline in Section 3.1, suppose that the two terms X_1 and X_2 have sample correlation zero. This means that, if $x_{ij}, i = 1, \dots, n$ and $j = 1, 2$ are the observed values of these two terms for the n cases in the data, $\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$.

3.4.1. Give the formula for the slope of the regression for Y on X_1, and for Y on X_2. Give the value of the slope of the regression for X_2 on X_1.

Solution: (1) $\hat{\beta}_1 = SX_1Y/SX_1X_1$; (2) $\hat{\beta}_2 = SX_2Y/SX_2X_2$; (3) $\hat{\beta}_3 = 0$. ∎

3.4.2. Give formulas for the residuals for the regressions of Y on X_1 and for X_2 on X_1. The plot of these two sets of residuals corresponds to the added-variable plot in Figure 3.1d.

Solution: (1) $\hat{e}_{1i} = y_i - \bar{y} - \hat{\beta}_1(x_{i1} - \bar{x}_1)$; (2) $\hat{e}_{3i} = x_{i2} - \bar{x}_2$. ∎

3.4.3. Compute the slope of the regression corresponding to the added-variable plot for the regression of Y on X_2 after X_1, and show that this slope is exactly the same as the slope for the simple regression of Y on X_2 ignoring X_1. Also find the intercept for the added variable plot.

Solution: Because $\sum \hat{e}_{3i} = 0$,

$$\begin{aligned}
\text{Slope} &= \sum \hat{e}_{3i}\hat{e}_{1i} / \sum \hat{e}_{3i}^2 \\
&= \sum (x_{i2} - \bar{x}_2)(y_i - \bar{y} - \hat{\beta}_1(x_{i1} - \bar{x}_1)) / \sum (x_{i2} - \bar{x}_2)^2 \\
&= \left(SX_2Y - \hat{\beta}_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \right) / SX_2X_2 \\
&= SX_2Y / SX_2X_2 \\
&= \hat{\beta}_2
\end{aligned}$$

The estimated intercept is exactly zero, and the R^2 from this regression is exactly the same as the R^2 from the regression of Y on X_2. ∎

4.1 Fit the regression of *Soma* on *AVE*, *LIN* and *QUAD* as defined in Section 4.1 for the girls in the Berkeley Guidance Study data, and compare to the results in Section 4.1.

Solution:

```
> summary(m1)    Mean function 1 from Table 4.1
Call:
lm(formula = Soma ~ WT2 + WT9 + WT18)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5921     0.6742    2.36   0.0212
WT2          -0.1156     0.0617   -1.87   0.0653
WT9           0.0562     0.0201    2.80   0.0068
WT18          0.0483     0.0106    4.56  2.3e-05

Residual standard error: 0.543 on 66 degrees of freedom
Multiple R-Squared: 0.566
F-statistic: 28.7 on 3 and 66 DF,  p-value: 5.5e-12

> summary(m2)     Mean function with transformed terms

Call:
lm(formula = Soma ~ AVE + LIN + QUAD)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4030 -0.2608 -0.0318  0.3801  1.4409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5921     0.6742    2.36   0.0212
AVE          -0.0111     0.0519   -0.21   0.8321
LIN          -0.0820     0.0304   -2.70   0.0089
QUAD         -0.0300     0.0162   -1.85   0.0688

Residual standard error: 0.543 on 66 degrees of freedom
Multiple R-Squared: 0.566,      Adjusted R-squared: 0.546
F-statistic: 28.7 on 3 and 66 DF,  p-value: 5.5e-12
```

(1) All summary statistics are identical. (2) All residuals are identical. (3) Intercepts are the same. The mean function for the first model is

$$\mathrm{E}(\mathit{Soma}|W) = \beta_0 + \beta_1 \mathit{WT2} + \beta_2 \mathit{WT9} + \beta_3 \mathit{WT18}$$

Substituting the definitions of *AVE*, *LIN* and *QUAD*, the mean function for the second model is

$$\begin{aligned}
\mathrm{E}(\mathit{Soma}|W) &= \eta_0 + \eta_1 \mathit{AVE} + \eta_2 \mathit{LIN} + \eta_3 \mathit{QUAD} \\
&= \eta_0 + \eta_1(\mathit{WT2} + \mathit{WT9} + \mathit{WT18})/3 \\
&\quad + \eta_2(\mathit{WT2} - \mathit{WT18}) + \eta_3(\mathit{WT2} - 2\,\mathit{WT9} + \mathit{WT18}) \\
&= \eta_0 + (\eta_1/3 + \eta_2 + \eta_3)\,\mathit{WT2} + (\eta_1/3 - 2\eta_3)\,\mathit{WT9} \\
&\quad + (\eta_1/3 - \eta_2 + \eta_3)\,\mathit{WT18}
\end{aligned}$$

which shows the relationships between the βs and the ηs (for example, $\hat{\beta}_1 = \hat{\eta}_1/3 + \hat{\eta}_2 + \hat{\eta}_3$). The interpretation in the transformed scale may be a bit easier, as only the linear trend has a small p-value, so we might be willing to describe the change in *Soma* over time as increasing by the same amount each year. ∎

4.2

4.2.1. Starting with (4.10), we can write

$$y_i = \mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x) + \varepsilon_i$$

Ignoring the error term ε_i, solve this equation for x_i as a function of y_i and the parameters.

Solution:

$$x_i = \mu_x + \frac{1}{\rho_{xy}}\frac{\sigma_x}{\sigma_y}(y_i - \mu_y)$$

This is undefined if $\rho_{xy} = 0$. ∎

4.2.2. Find the conditional distribution of $x_i|y_i$. Under what conditions is the equation you obtained in Problem 4.2.1, which is computed by inverting the regression of y on x, is the same as the regression of x on y?

Solution: Simply reverse the role of x and y in (4.10) to get

$$x_i|y_i \sim \mathrm{N}\left(\mu_x + \rho_{xy}\frac{\sigma_x}{\sigma_y}(y_i - \mu_y), \sigma_y^2(1 - \rho_{xy}^2)\right)$$

These two are the same if and only if the correlation is equal to plus or minus one. In general there are two regressions. ∎