Module 4 - SLR inference

STAT 401

Section 1: Inference for OLS estimators

We found last week that under the standard assumptions of the SLR model, (including $e_i \sim N(0, \sigma^2)$ for all i),

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right)$$

Equivalently,

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}} \sim N(0,1)$$

and so

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{SXX}} \sim t_{n-2}$$

We can therefore say that

$$P\left(t_{\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{SXX}} < t_{1-\alpha/2}\right) = 1 - \alpha.$$

Rearranging inside the probability expression, we get

$$P\left(\hat{\beta}_1 - t_{1-\alpha/2}\hat{\sigma}/\sqrt{SXX} < \beta_1 < \hat{\beta}_1 - t_{\alpha/2}\hat{\sigma}/\sqrt{SXX}\right) = 1 - \alpha.$$

Thus, a $(1-\alpha)*100\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{SXX}}$$

Likewise, a $(1-\alpha)*100\%$ confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{\alpha/2}\hat{\sigma}\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)$$

For example, in the spring hardness data:

$$\hat{\beta}_1 = -1.266, \quad SXX = 1950, \quad \hat{\sigma}^2 = 2.235, \quad n = 14$$

For a 90% confidence interval for β_1, $t_{0.05,12} = -1.782$ and the interval is given by

$$-1.266 \pm 1.782 * \sqrt{\frac{2.235}{1950}} = (-1.206, -1.326)$$

A 90% confidence interval for β_0 requires the extra information

$$\bar{x} = 45$$

The resulting interval is

$$94.134 \pm 1.782 * \sqrt{2.235\left(\frac{1}{14} + \frac{45^2}{1950}\right)} = (91.33, 96.94)$$

The interpretation of this last interval is:

> *We are 90% confident that the average hardness for all springs in this population which are quenched in a bath temperature of 0 is between 91.33 and 96.94.*

The interpretation of the former interval is:

> *We are 90% confident that for every additional degree of quench bath temperature, the average hardness increases between -1.206 and -1.326 units.*

R can produce these intervals:

```
Model <- lm(Hard ~ Temp, data = springs)
confint(Model, level = 0.9)

##                   5 %      95 %
## (Intercept) 91.326952 96.941180
## Temp        -1.326493 -1.205814
```

Hypothesis tests of the slope β_1 commonly presume a null hypothesis of no linear association between X and Y. This is accomplished via the hypothesis: $H_0 : \beta_1 = 0$. One- and two-sided hypotheses are testable. Under this null hypothesis, the quantity

$$t = \frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)}$$

has a t distribution with $n - 2$ df. The corresponding p-value is the area under the density curve for this distribution in the appropriate tail(s).

Hypotheses that test a claim about β_0 are less common, but still possible. Under the null hypothesis that

$$H_0 : \beta_0 = \beta_{0_0}$$

the test statistic

$$t = \frac{\beta_0 - \beta_{0_0}}{S.E.(\hat{\beta}_0)}$$

has a t distribution with $n - 2$ df.

If we desired to test for a linear association between `Hard` and `Temp`, we would test the null hypothesis:

$$H_0 : \beta_1 = 0$$

We will test the two-sided alternative. The test statistic is

$$t = \frac{-1.266}{\sqrt{\frac{2.235}{1950}}} = -37.40$$

resulting in a p-value of 8.6×10^{-14}. At $\alpha = 0.05$, we would to reject the null hypothesis.

Our interpretation of the test would be

> *At a level 0.05 test, there is statistically significant evidence that there is a linear assocation between quench bath temperature and spring hardness.*

If we desired to test a hypothesis about β_0, such as

$$H_0 : \beta_0 = 90$$

versus the left-sided alternative ($H_A : \beta_0 < 90$), the test statistic is

$$t = \frac{94.134 - 90}{\sqrt{2.235\left(\frac{1}{14} + \frac{45^2}{1950}\right)}} = 2.62$$

The resulting p-value is 0.99. At $\alpha = 0.01$, we would fail to reject the null hypothesis.

Our interpretation of the test would be

> *At a level 0.01 test, there is not statistically significant evidence that the mean spring hardness of all springs quenched in a bath of temperature 0 is less than 90.*

R can produce these such tests:

```
summary(Model)
```

```
##
## Call:
## lm(formula = Hard ~ Temp, data = springs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5494 -1.1898 -0.3687  0.5986  2.9505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 94.13407    1.57501   59.77 3.18e-16 ***
## Temp        -1.26615    0.03386  -37.40 8.58e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.495 on 12 degrees of freedom
## Multiple R-squared:  0.9915,^^IAdjusted R-squared:  0.9908
## F-statistic:  1399 on 1 and 12 DF,  p-value: 8.578e-14
```

Section 2: Inference for functions of O.L.S. estimators

A $(1-\alpha)*100\%$ confidence interval for the mean response at a particular value of *x* is given by

$$\hat{y} \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SXX}}$$

A $(1-\alpha)*100\%$ prediction interval for a new response at a particular value of *x* is given by

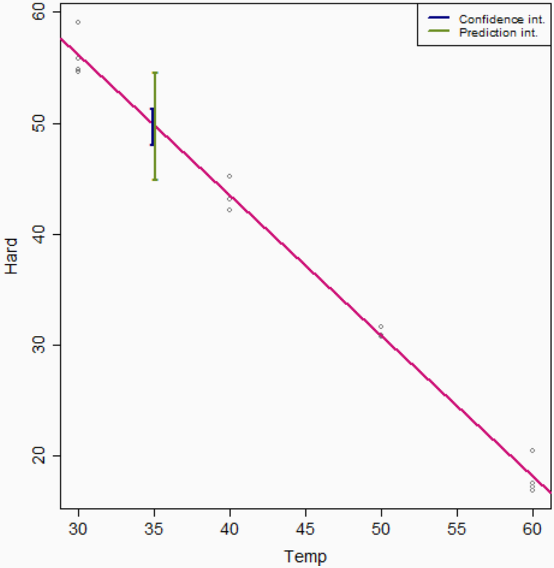
$$\hat{y} \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SXX}}$$

For example, in the spring hardness data, suppose the engineers wish to know more about spring hardness when quench bath temperature is 35. A 99% confidence interval for the mean hardness of all such springs is given by

$$49.824 \pm 3.055 * 1.495 \sqrt{\frac{1}{14} + \frac{(35-45)^2}{1950}} = (48.22, 51.42)$$

A 99% prediction interval for the hardness of a newly measured spring cooled at 35 degrees is given by

$$49.824 \pm 3.055 * 1.495 \sqrt{1 + \frac{1}{14} + \frac{(35 - 45)^2}{1950}} = (44.98, 54.66)$$


Confidence int.
Prediction int.
Hard
60
50
40
30
20
Temp
30
35
40
45
50
55
60

R can produce these intervals:

```r
predict(Model, newdata = data.frame(Temp = 35),
    interval = "confidence", level = 0.99)
```

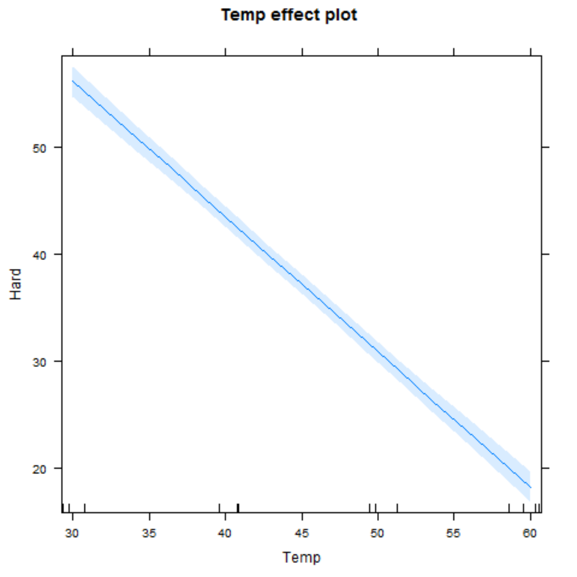
```
##        fit      lwr      upr
## 1 49.81868 48.21902 51.41835
```

```r
predict(Model, newdata = data.frame(Temp = 35),
    interval = "prediction", level = 0.99)
```

```
##        fit      lwr      upr
## 1 49.81868 44.98006 54.65731
```

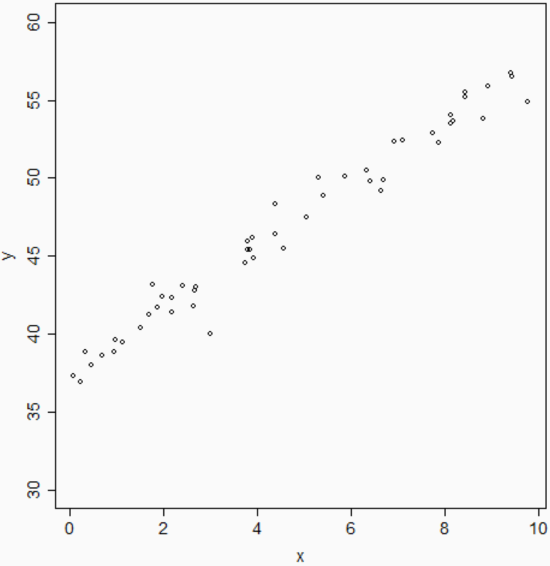
Section 3: Effect plots

Effect plots are a simple way to visualize the effect of a predictor on the response. In the S.L.R. setting, an effect plot is simply the plot of the fitted regression line, but in more complex settings these plots become more insightful. We examine the effect plot of the springs hardness model:


Temp effect plot
Hard
50
40
30
20
30
35
40
45
50
55
60
Temp

Section 4: ANOVA tables

We have developed a test for β_1 and discussed r_{xy} as ways of measuring the strength of a linear regression relationship. There is another (closely-related) way to measure this relationship. To develop it, we use with a data set which displays good linear association.


60
55
50
45
40
35
30
y
0
2
4
6
8
10
x

How can predictions for new draws of Y be made? A naive prediction is $\bar{y}$. In other words, a possible (but bad) model for this data is

$$\hat{E}(Y) = \hat{\beta}_0 = \bar{y}$$

since $\hat{\beta}_0 = \bar{y}$ and β_1 is assumed 0. This model is assuming that we can predict the response just as well without the predictor as we can with it.

The ANOVA approach focuses on variability. What is $\hat{\sigma}^2$ in this case?

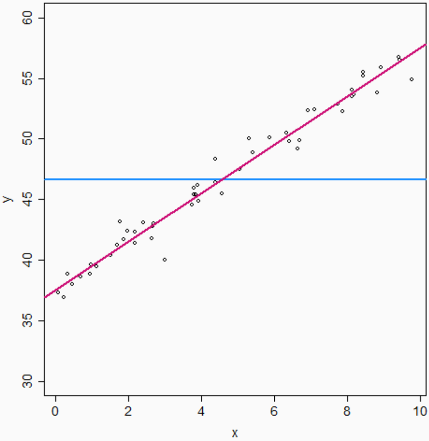
$$\begin{aligned}
\hat{\sigma}^2 &= \frac{RSS}{n-2} \\
&= \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} \\
&= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2} \\
&= \frac{SYY}{n-2} \\
&\doteq s_y^2
\end{aligned}$$

This is an extreme case that we want to use as a benchmark. We might refer to *SYY* as the *Sum of Squares Total*:

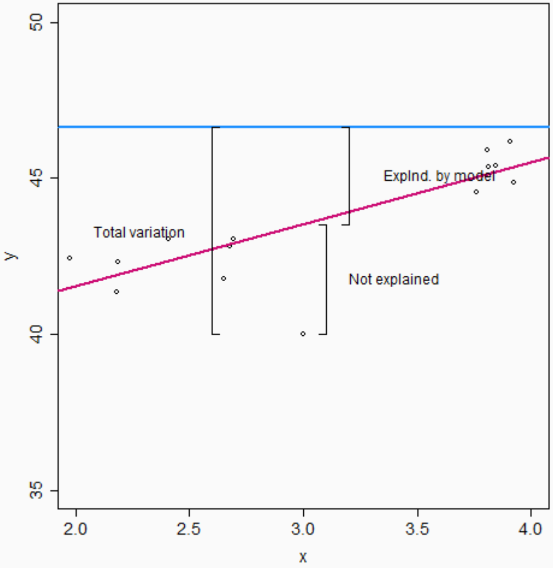
$$SYY = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

because it represents the total amount of variability in *y* when no predictors are included.

Clearly there is a better model available here than $\hat{E}(Y) = \bar{y}$:



What does it mean to say that $\hat{E}(Y) = \hat{\beta}_0 + \hat{\beta}_1 x$ fits better than $\hat{E}(Y) = \hat{\beta}_0 = \bar{y}$ in this example? One way to explain this claim is to point out that, for most points, the distance from the point to the best-fit line ($\hat{e}_i$) is small and the distance from the best-fit line to the mean of y is big.


50
45
40
35
y
2.0
2.5
3.0
3.5
4.0
x
Total variation
Explnd. by model
Not explained

In math symbols, for any i in $1, \ldots, n$,

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y} - \bar{y})$$

and for this data, $(\hat{y} - \bar{y})$ is large relative to $(y_i - \hat{y})$. In contrast, in the regression of y on x, $(\hat{y} - \bar{y})$ was small relative to $(y_i - \hat{y})$.

We take the identity above and square both sides to eliminate negative signs:

$$(y_i - \bar{y})^2 = ((y_i - \hat{y}) + (\hat{y} - \bar{y}))^2$$

and we sum up over the *n* data points:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}((y_i - \hat{y}) + (\hat{y} - \bar{y}))^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y})^2 + \sum_{i=1}^{n}(\hat{y} - \bar{y})^2\end{aligned}$$

This identity says that the total amount of variability in the data set, *SYY*, equals the total amount of squared residuals, *RSS*, plus something else. The something else is the total amount of squared distance from the mean to the best-fit line. In other words,

$$\begin{aligned}\text{Total variability} = \text{Variability explained by model} + \\ \text{Variability not explained by model}\end{aligned}$$

This is the essence of ANOVA: decomposing the total variability in the response into a part which is explained by the regression model and a part which is not.

The variability explained by the model is termed *SSreg*, or *Sum of squares for regression*. Therefore,

$$SYY = SSreg + RSS$$

All sums of squares have degrees of freedom. Informally, the degrees of freedom respresent how many independent pieces of information are being summed. We already know that the *df* for *RSS* is $n - 2$. For

$$SYY = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

the *df* is $n - 1$. For

$$SSreg = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$

the *df* is 1.

The size of any sum of squares depends not just on the amount of variability but on its degrees of freedom. Distinct sums of squares cannot be compared for this reason. But if we divide each sum of squares by its degrees of freedom, the resulting *mean squares* can be compared.

$$MSreg = \frac{SSreg}{1}$$

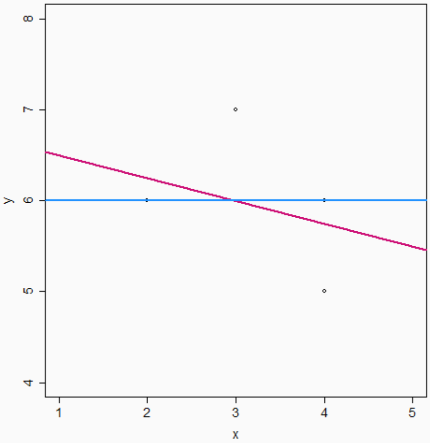
$$RMS = \frac{RSS}{n-2}$$

Note that $RMS = \hat{\sigma}^2$.

The quantities *SYY*, *SSreg*, *RSS*, *MSreg*, and *RMS* are customarily displayed in an ANOVA table:

Source	df	SS	MS
Model	1	*SSreg*	*MSreg*
Residual	$n-2$	*RSS*	*RMS*
Total	$n-1$	*SYY*	

For example, return to the disease remission data.



In this data, we found that $\bar{y} = 6$ and

x	y	$\hat{y}$
4	6	5.75
2	6	6.25
3	7	6
4	5	5.75
2	6	6.25

$$SYY =$$

$$(6-6)^2 + (6-6)^2 + (7-6)^2 + (5-6)^2 + (6-6)^2 = 2$$

$$SSreg =$$

$$(5.75-6)^2 + (6.25-6)^2 + (5.75-6)^2 + (6.25-6)^2 = 1/4$$

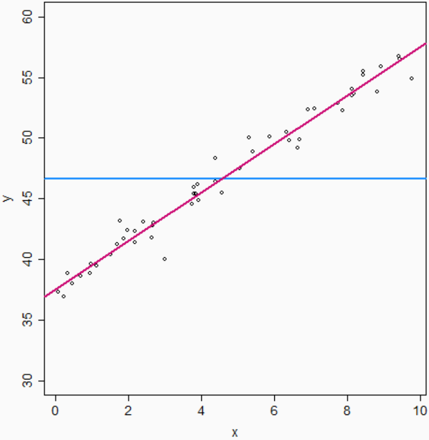
$$RSS =$$

$$(6-5.75)^2 + (6-6.25)^2 + 1 + (5-5.75)^2 + (6-6.25)^2 = 7/4$$

The ANOVA table is:

Source	df	SS	MS
Model	1	1/4	1/4
Residual	3	7/4	7/12
Total	4	2	

In contrast, for the example data with *y* and *x* from earlier,



```
anova(lm(y ~ x, data = Linear))

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 1703.45 1703.45  1668.7 < 2.2e-16 ***
## Residuals 49   50.02    1.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Section 5: R^2

If SYY represents the total amount of (squared) variability of y about its mean, and RSS represents the total amount of (squared) variability of y about the best-fit line, then the ratio

$$\frac{RSS}{SYY}$$

represents the percentage of total (squared) variability not accounted for by the best-fit line. And

$$1 - \frac{RSS}{SYY} = \frac{SSreg}{SYY} = R^2$$

In other words, R^2 measures the strength of the linear regression fit. R^2 is known as the *coefficient of determination* and

$$0 \leq R^2 \leq 1$$

where values closer to 1 indicate a stronger fit.

The textbook shows that

$$SSreg = \frac{SXY^2}{SXX}$$

which allows us to deduce that

$$R^2 = \frac{SSreg}{SYY} = \frac{SXY^2/SXX}{SYY} = \frac{SXY^2}{SXX \cdot SYY} = r_{xy}^2$$

so that the coefficient determination is the correlation coefficient squared.

In the disease remission data,

$$R^2 = \frac{SSreg}{SYY} = \frac{0.25}{2} = 0.125$$

The interpretation of this would be:

12.5% of the variability in y is explained by the linear regression on x

In contrast, for the example data with *y* and *x* from earlier,

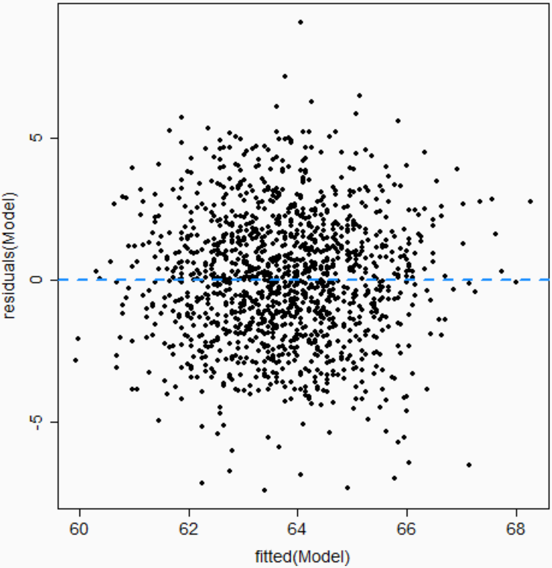
```
summary(lm(y ~ x, data = Linear))

## 
## Call:
## lm(formula = y ~ x, data = Linear)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5228 -0.4973  0.1202  0.6021  2.0380 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 37.55051    0.26386  142.31   <2e-16 ***
## x            1.99076    0.04873   40.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.01 on 49 degrees of freedom
## Multiple R-squared:  0.9715,^^IAdjusted R-squared:  0.9709 
## F-statistic:  1669 on 1 and 49 DF,  p-value: < 2.2e-16
```

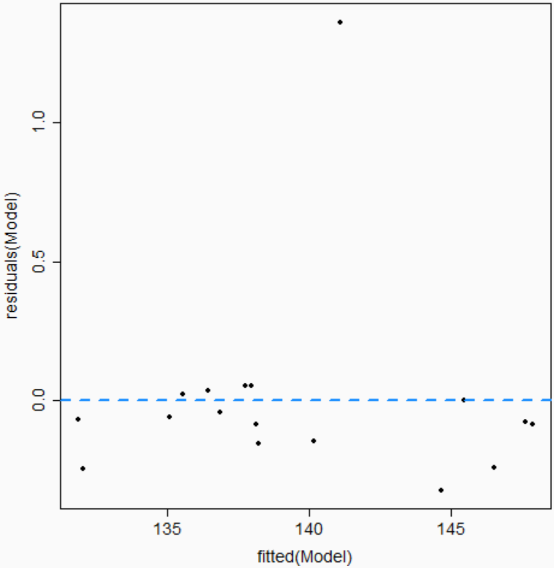
Section 6: Residuals

Residuals provide clues about whether the model's assumptions are essentially correct or not. Recall residuals are defined as

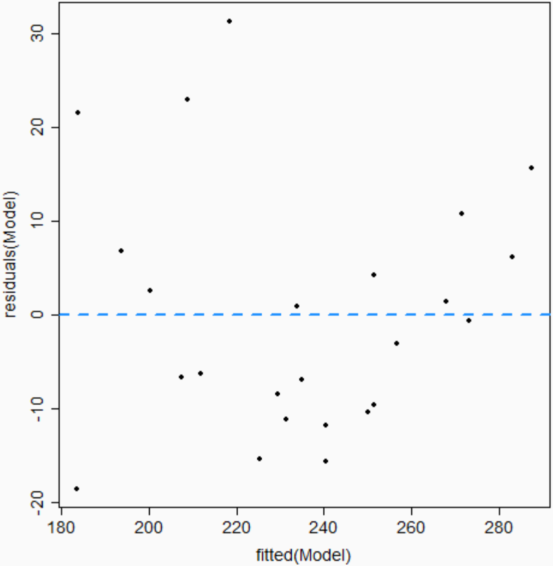
$$\hat{e}_i = y_i - \hat{y} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

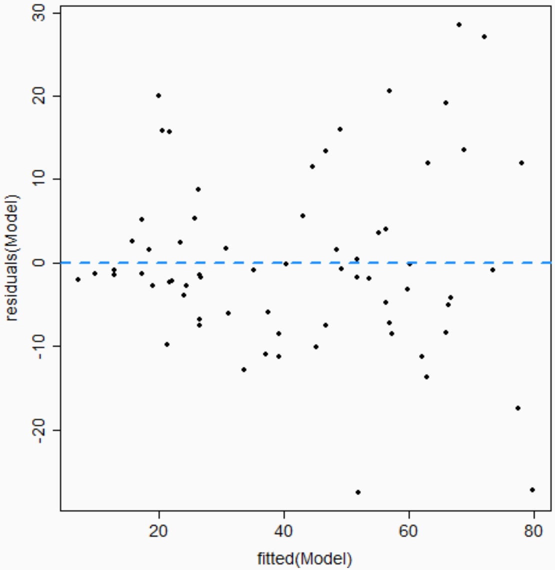
and can be thought of as leftover variation in Y that the model did not explain. A plot of these distances against the fitted values $\hat{y}_i$ provides insight into the nature of this unexplained variation.


residuals(Model)
fitted(Model)
5
0
-5
60
62
64
66
68



residuals(Model)
1.0
0.5
0.0
135
140
145
fitted(Model)


residuals(Model)
fitted(Model)
-20
-10
0
10
20
30
180
200
220
240
260
280



30
20
10
0
-10
-20
residuals(Model)
20
40
60
80
fitted(Model)

When a point is outlying or influential as in the Forbes data, it can be instructive to fit the model both with and without the point.

```
summary(lm(lpres ~ bp, data = Forbes))
```

```
##
## Call:
## lm(formula = lpres ~ bp, data = Forbes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32220 -0.14473 -0.06664  0.02184  1.35978
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.13778    3.34020  -12.62 2.18e-09 ***
## bp            0.89549    0.01645   54.43  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.379 on 15 degrees of freedom
## Multiple R-squared:  0.995,^^IAdjusted R-squared:  0.9946
## F-statistic:  2963 on 1 and 15 DF,  p-value: < 2.2e-16
```

```
summary(lm(lpres ~ bp, data = Forbes,
    subset = -12))
```

```
##
## Call:
## lm(formula = lpres ~ bp, data = Forbes, subset = -12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21129 -0.06132  0.01627  0.09152  0.13110
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41.30838    1.00052  -41.29    5e-16 ***
## bp            0.89099    0.00493  180.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1133 on 14 degrees of freedom
## Multiple R-squared:  0.9996,^^IAdjusted R-squared:  0.9995
## F-statistic: 3.266e+04 on 1 and 14 DF,  p-value: < 2.2e-16
```

Section 7: Deviations from the mean

At times it can be convenient to center (and/or scale) a predictor's data before using it in a regression model. This might be for reasons such as:

1. Standardized coefficients
2. Numerical stability of the model-fitting
3. Model interpretability

Start with the basic S.L.R. model and then center x_i:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + e_i \\ &= \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + e_i \\ &= \beta_0 + \beta_1\bar{x} + \beta_1(x_i - \bar{x}) + e_i \\ &= \alpha + \beta_1 x_i^* + e_i \end{aligned}$$

where $x_i^* = (x_i - \bar{x})$ and $\alpha = \beta_0 + \beta_1\bar{x}$.

The interpretation of β_1 is unchanged, but the new interpretation of α is

> *When the predictor is set at its mean, the mean response is [alpha].*

Finding least squares estimators can be done by minimizing the function $RSS(a, b_1) = \sum_{i=1}^{n}(y_i - a - b_1 x_i^*)^2$. Alternatively, see that

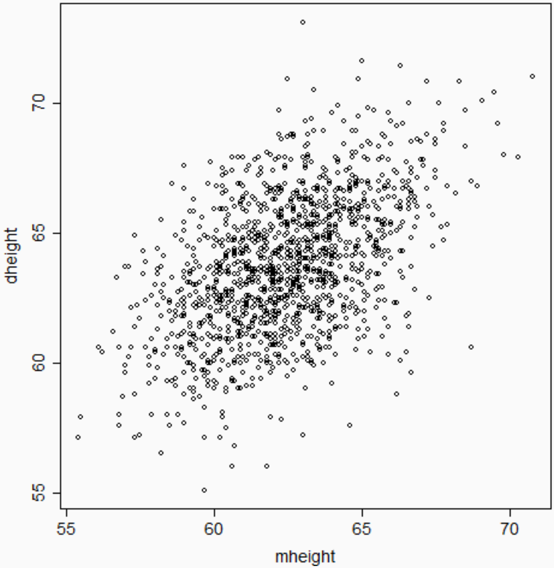
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \text{ and } \alpha = \beta_0 + \beta_1\bar{x}$$

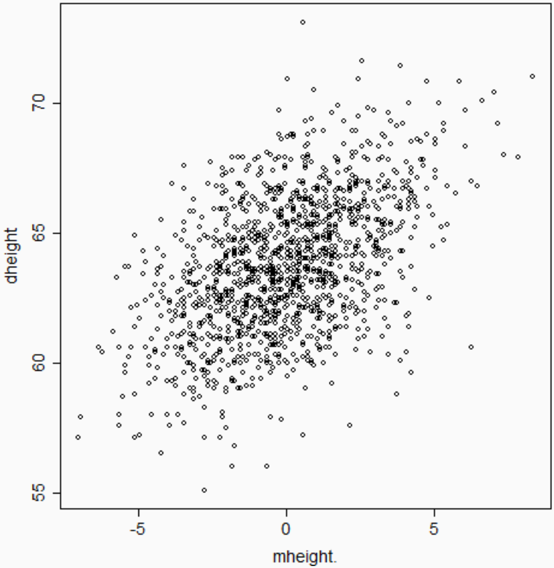
implies that

$$\begin{aligned}\hat{\alpha} &= \hat{\beta}_0 + \hat{\beta}_1\bar{x} \\ &= \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} \\ &= \bar{y}\end{aligned}$$

What would be the variance of $\hat{\alpha}$ and the covariance of $\hat{\alpha}$ and $\hat{\beta}_1$?

$$
\begin{aligned}
V(\hat{\alpha}) &= V(\bar{Y}) \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$


55
60
65
70
dheight
55
60
65
70
mheight


dheight
70
65
60
55
-5
0
5
mheight.

```
summary(lm(dheight ~ mheight, data = Heights))


##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408,^^IAdjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```
Heightsmheight. <- scale(Heightsmheight,
    scale = FALSE)
summary(lm(dheight ~ mheight., data = Heights))
```

```
##
## Call:
## lm(formula = dheight ~ mheight., data = Heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 63.75105    0.06112 1043.08   <2e-16 ***
## mheight.     0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408,^^IAdjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

The interpretation of $\hat{\alpha}$ in this data:

When the mother's height is 62.5 inches, the mean response is 63.8 inches.