STAT 135 Lab 12
Multiple Linear Regression (Matrix Form), Residual analysis, Inference about $\hat{\beta}$

Rebecca Barter

April 20, 2015

Multiple Linear Regression

Simple Linear Regression

Recall simple linear regression, where we had a single variable, x, and we wanted to use it as a predictor for our response, y:

$$y = \beta_0 + \beta_1 x + \epsilon$$

So that taking a linear combination of x (with some noise) gives the response y.
For example, we could fit a model for the height of suds (mm) as a function of soap (g) used based on the following 10 observations.

i	1	2	3	4	5	6	7	8	9	10
soap (x)	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
suds (y)	24.4	32.1	37.1	40.4	43.3	51.4	61.9	66.1	77.2	79.2

Simple Linear Regression

For example, we could fit a model for the height of suds (mm) as a function of soap (g) used based on the following 10 observations.

i	1	2	3	4	5	6	7	8	9	10
soap (x)	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
suds (y)	24.4	32.1	37.1	40.4	43.3	51.4	61.9	66.1	77.2	79.2

and find that

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = -20.2$$

$$\hat{\beta}_1 = cov(x, y)/var(y) = 12.4$$

and conclude that our fitted line is given by:

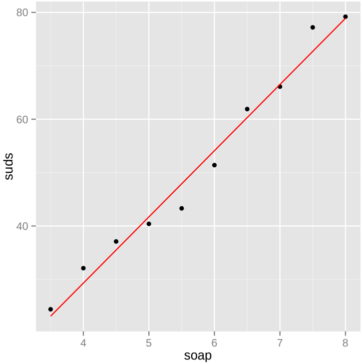
$$\widehat{\text{suds}} = -20.2 + 12.4 \times \text{soap}$$

I

Simple Linear Regression

Our fitted (red) line is given by:

$$\widehat{\text{suds}} = -20.2 + 12.4 \times \text{soap}$$



Multiple Linear Regression

What if we believed that not only the amount of soap used, but also the following variables affected the height of the suds:

- oil level of soap
- acidity of the soap
- amount of water
- hardness of the water

Then we might want to include these predictors in our model:

$$\begin{aligned}\text{suds} \ = \ \beta_0 \ + \ \beta_1 \times (\text{soap}) \ + \ \beta_2 \times (\text{oil}) \ + \ \beta_3 \times (\text{acidity}) \\ + \ \beta_4 \times (\text{water}) \ + \ \beta_5 \times (\text{hardness}) + \epsilon\end{aligned}$$

This is an example of **multiple linear regression*** (linear regression with more than one *predictor* (x)).

*not the same as **multivariate linear regression**: lin. reg. with more than one *response* (y)

Multiple Linear Regression

In multiple linear regression, the **response, $\mathbf{y_i}$**, for subject i ($i = 1, ..., n$) can be modeled as a linear combination of your $p - 1$ **predictors**, $\mathbf{x_{i,j}}$, $j = 0, ..., p - 1$, as follows:

$$\boxed{y_i = \beta_0 \ + \ \beta_1 x_{i,1} \ + \ \beta_2 x_{i,2} \ + ... + \ \beta_{p-1} x_{i,(p-1)} \ + \ \epsilon_i}$$

where the **random error terms**, $\boldsymbol{\epsilon}_i$, are independent random variables such that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

- Note that we often require that $\epsilon_i \overset{IID}{\sim} N(0, \sigma^2)$, but in general, the normality is not required for unbiased estimates of the β's
- We do require $p < n$ (i.e. the number of predictors is less than the number of observations)

Multiple Linear Regression

Our regression equation is of the form:

$$y_i = \beta_0 \ + \ \beta_1 x_{i,1} \ + \ \beta_2 x_{i,2} \ + ... + \ \beta_{p-1} x_{i,(p-1)} \ + \ \epsilon_i$$

The question is, how do we estimate the β's? We want to minimize the distance from our observed observations y_i to our fitted line. That is, we want to find the values of β_0, β_1, ..., β_{p-1} that minimize

$$S(\beta_0, \beta_1, ..., \beta_{p-1}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - ... - \beta_{p-1} x_{i,p-1})^2$$

To do this by differentiating with respect to each β_j separately will be very time consuming.

Multiple Linear Regression

We want to find the values of β_0, β_1, ..., β_{p-1} that minimize

$$S(\beta_0, \beta_1, ..., \beta_{p-1}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - ... - \beta_{p-1} x_{i,p-1})^2$$

At this point, it's time to introduce the much more convenient **matrix notation** for linear regression.

Matrix Notation for Multiple Linear Regression

Multiple Linear Regression

Our regression equation for each subject ($i = 1, ..., n$) are of the form:

subject 1: $y_1 = \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + ... + \beta_{p-1} x_{1,(p-1)} + \epsilon_1$

subject 2: $y_2 = \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + ... + \beta_{p-1} x_{2,(p-1)} + \epsilon_2$

$$\vdots$$

subject n: $y_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + ... + \beta_{p-1} x_{n,(p-1)} + \epsilon_n$

Multiple Linear Regression

subject 1: $y_1 = \beta_0 \times 1 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + ... + \beta_{p-1} x_{1,(p-1)} + \epsilon_1$

subject 2: $y_2 = \beta_0 \times 1 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + ... + \beta_{p-1} x_{2,(p-1)} + \epsilon_2$

$$\vdots$$

subject n: $y_n = \beta_0 \times 1 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + ... + \beta_{p-1} x_{n,(p-1)} + \epsilon_n$

This can be written in matrix form as follows

$$\underset{(n\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} = \underset{(n\times p)}{\begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{bmatrix}} \underset{(p\times 1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}} + \underset{(n\times 1)}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

Multiple Linear Regression

Multiple regression equations in matrix form:

$$Y = X\beta + \epsilon$$

$$\underset{(n\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} = \underset{(n\times p)}{\begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{bmatrix}} \underset{(p\times 1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}} + \underset{(n\times 1)}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

- Y is the $(n \times 1)$ (observed) **response vector**
- X is the $(n \times p)$ (observed) **design matrix**
- β is the $(p \times 1)$ (unobserved) **coefficient vector**
- ϵ is the $(n \times 1)$ (unobserved) **error vector**

Multiple Linear Regression

$$Y = X\beta + \epsilon$$

and we have the following assumptions on the error term:

$$E[\epsilon] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad Cov(\epsilon) = \sigma^2 I_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Moreover, we need ϵ to be independent of X.

Note that we still need to figure out how to estimate our β's

Multiple Linear Regression

Note that finding the values of $\beta_0, \beta_1, ...\beta_{p-1}$ that minimizes

$$S(\beta_0, \beta_1, ..., \beta_{p-1}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - ... - \beta_{p-1} x_{i,p-1})^2$$

is the same as finding the vector $\boldsymbol{\beta}$ that minimizes

$$\boxed{S(\boldsymbol{\beta}) = \|Y - X\boldsymbol{\beta}\|_2^2 = (Y - X\boldsymbol{\beta})^T (Y - X\boldsymbol{\beta})}$$

(this is just a number: $(Y - X\boldsymbol{\beta})^T (Y - X\boldsymbol{\beta})$ has dim (1×1))

---

where $\| \cdot \|_2$ is the L^2-norm which is defined for a vector $\mathbf{x} = (x_1, ..., x_n)$ by

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^T \mathbf{x}}$$

Multiple Linear Regression

We want to find the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$ that minimizes

$$S(\boldsymbol{\beta}) = \|Y - X\boldsymbol{\beta}\|_2^2 = (Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta})$$

We will show that our optimal $\boldsymbol{\beta}$ is given by

$$\boxed{\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y}$$

Where $\widehat{\boldsymbol{\beta}}$ is estimated vector of coefficients for our predictors in our fitted regression line

Multiple Linear Regression

To show that $\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y$, we want to find the vector $\boldsymbol{\beta}$ that satisfies

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

where

$$\begin{aligned}
\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (Y - X\boldsymbol{\beta})^T (Y - X\boldsymbol{\beta}) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left[Y^T Y - \boldsymbol{\beta}^T X^T Y - Y^T X \boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} \right] \\
&= -X^T Y - X^T Y + 2 X^T X \boldsymbol{\beta}
\end{aligned}$$

---

where we have used the following differentiation identities ($\mathbf{x}$ and $\mathbf{a}$ are both $(p \times 1)$ vectors and A is a symmetric $(p \times p)$ matrix):

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^T \mathbf{x} = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{a} = \mathbf{a} \qquad \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^{\mathbf{T}} A \mathbf{x} = 2A\mathbf{x} = 2\mathbf{x}^T A$$

Multiple Linear Regression

Thus we need to solve for $\widehat{\boldsymbol{\beta}}$ in:

$$-X^T Y - X^T Y + 2X^T X\widehat{\boldsymbol{\beta}} = 0$$

rearranging gives the **normal equations**

$$X^T X\widehat{\boldsymbol{\beta}} = X^T Y$$

so that if $(X^T X)$ is invertible gives

$$\boxed{\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y}$$

and thus our **fitted model** is given by

$$\boxed{\widehat{Y} = X\widehat{\boldsymbol{\beta}} = X\left(X^T X\right)^{-1} X^T Y}$$

Exercise

Exercise: Multiple Linear Regression

Show that when we have simple linear regression ($p = 2$), i.e.

$$\underset{(n\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} = \underset{(n\times 2)}{\begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \\ 1 & x_{3,1} \\ \vdots & \vdots \\ 1 & x_{n,1} \end{bmatrix}} \underset{(2\times 1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}} + \underset{(n\times 1)}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

that

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y$$

yields the same estimates as when we solved for β_0 and β_1 individually.

Exercise

Exercise: Multiple Linear Regression

The file cigarete_dat.txt contains measurements of weight and tar, nicotine, and carbon monoxide content for 25 brands of domestic cigarettes. We want to use this dataset to define a model for the carbon monoxide level as a function of the other variables.

- Identify the correlation between all pairs of variables. If we find two variables are highly correlated, should we use only one of them?
- Choose the variables to include in the linear model, and identify the design matrix, X.
- Fit a multiple linear regression model to CO levels and report the fitted model.

Inference about $\widehat{\boldsymbol{\beta}}$

Inference about $\widehat{\boldsymbol{\beta}}$

We now turn to examining properties of our estimator (a random vector), $\hat{\boldsymbol{\beta}}$, in particular, calculating its expected value and covariance matrix. Recall that

- X and $\boldsymbol{\beta}$ are fixed
- $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is random (but only through epsilon)
- $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2 I_{n\times n}$

$$\begin{aligned}
E\left(\widehat{\boldsymbol{\beta}}\right) &= E\left(\left(X^T X\right)^{-1} X^T Y\right) \\
&= \left(X^T X\right)^{-1} X^T E\left(Y\right) \\
&= \left(X^T X\right)^{-1} X^T E\left(X\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) \\
&= \left(X^T X\right)^{-1} X^T X\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

Thus $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$

Inference about $\widehat{\boldsymbol{\beta}}$

Recall that

- X and $\boldsymbol{\beta}$ are fixed
- $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is random (but only through epsilon)
- $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2 I_{n\times n}$

$$
\begin{aligned}
Cov\left(\widehat{\boldsymbol{\beta}}\right) &= Cov\left(\left(X^T X\right)^{-1} X^T Y\right) \\
&= \left(X^T X\right)^{-1} X^T Cov(Y) X \left(X^T X\right)^{-1} \\
&= \left(X^T X\right)^{-1} X^T Cov(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) X \left(X^T X\right)^{-1} \\
&= \left(X^T X\right)^{-1} X^T Cov(\boldsymbol{\epsilon}) X \left(X^T X\right)^{-1} \\
&= \left(X^T X\right)^{-1} X^T \sigma^2 I_{n\times n} X \left(X^T X\right)^{-1} \\
&= \sigma^2 \left(X^T X\right)^{-1}
\end{aligned}
$$

---

where we have used the formula that if A is a constant matrix then $Cov(AY) = ACov(Y)A^T$ and $Cov(A+Y) = cov(A)$

Inference about $\widehat{\boldsymbol{\beta}}$

$$Cov\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 \left(X^T X\right)^{-1}$$

We don't know σ^2, but we can estimate it using:

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\|Y - X\hat{\beta}\|_2^2}{n-p} = \frac{\|e\|_2^2}{n-p}$$

where $e = Y - X\hat{\beta}$ is the vector of residuals.*

---

*See last week's slides.

Inference about $\widehat{\boldsymbol{\beta}}$

We can also test hypotheses about our true parameters, β_j. It is common to test the hypothesis that

$$H_0 : \beta_j = 0$$

for each $j = 0, .., p-1$, since if we have that $\beta_j = 0$, then the variable to which it corresponds has no effect on our response.

Note that if we assumed, $\epsilon \sim N(0, \sigma^2 I_{n\times n})$ then we have shown that

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

where we can estimate our variance by $\hat{\sigma}^2 (X^T X)^{-1}$

Inference about $\widehat{\boldsymbol{\beta}}$

$$H_0 : \beta_j = 0$$

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

So

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{k,k}}} \sim t_{n-p}$$

so our p-value is given by

$$P_{H_0}\left(|t_{n-p}| \geq \left| \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{k,k}}} \right| \right)$$

where $(X^T X)^{-1}_{k,k}$ is the kth diagnoal entry of the matrix $(X^T X)^{-1}$

Exercise

Exercise: Inference about $\widehat{\boldsymbol{\beta}}$

For our previous example, test the hypothesis that the coefficient (β_{tar}) of Tar is zero.

Model Diagnostics

Model Diagnostics

How can we tell if our model is a good fit for the data?

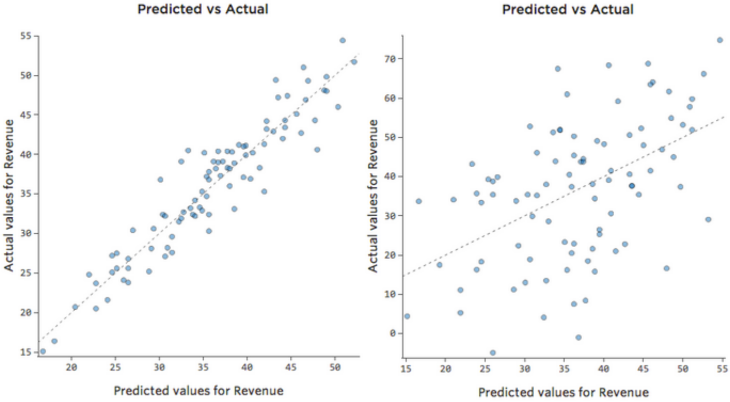
- Fitted-versus-observed plot
- Standardized residual versus fitted values plot

Model Diagnostics

We can look at a fitted-versus-observed plot

- We hope to see that the fitted values are very similar to the observed (true) values
- This corresponds to a diagonal line.

Model Diagnostics



The fitted/predicted values for the model on the left are closer to the true values than for the model on the right.

Model Diagnostics

We can look at a standardized residual versus fitted values plot

- We hope to see that the residuals are randomly scattered about $y = 0$
- We don't want to see any clear pattern

Note that the standardized residuals are

$$\frac{e_i}{\sqrt{Var(e_i)}}$$

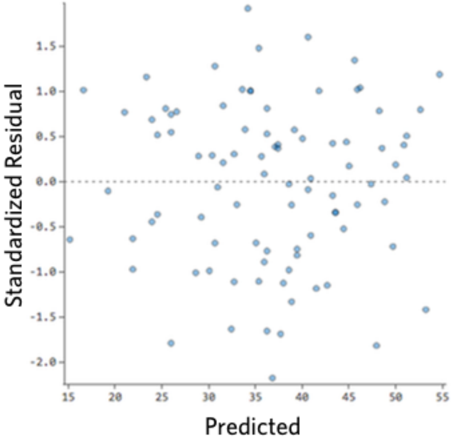
where

$$\begin{aligned}
Cov(e) &= Var(Y - X\hat{\beta}) \\
&= Var(Y - X(X^TX)^{-1}X^TY) \\
&= Var((I - X(X^TX)^{-1}X^T)Y) \\
&= (I - X(X^TX)^{-1}X^T)Var(Y)(I - X(X^TX)^{-1}X^T) \\
&= \sigma^2(I - X(X^TX)^{-1}X^T)
\end{aligned}$$

where the last equality follows from
$(I - X(X^TX)^{-1}X^T)^2 = (I - X(X^TX)^{-1}X^T)$

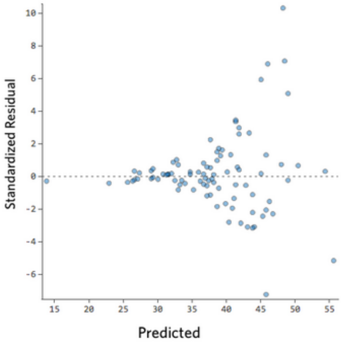
Model Diagnostics

This is an example of a good residual plot



Model Diagnostics

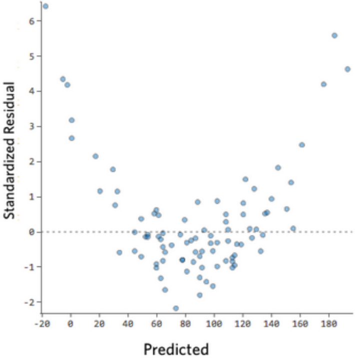
This is an example of a residual plot where there is heteroscedasticity (the variance is not the same for all observations)



A common solution is to transform a variable (most often a log-transform). Sometimes heteroscedasticity indicates that an important variable is missing.

Model Diagnostics

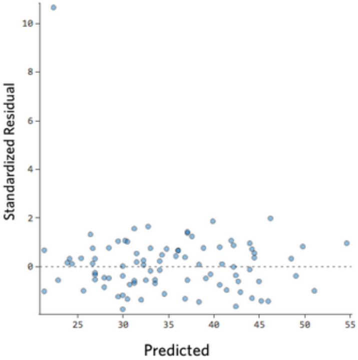
This is an example of a residual plot where there is a non linear relationship between variable and response



We might include a non-linear variable in our model. For example if the above model corresponded to $y = x + 1$, we might instead use $y = x^2 + x + 1$ (define a new predictor equal to x^2)

Model Diagnostics

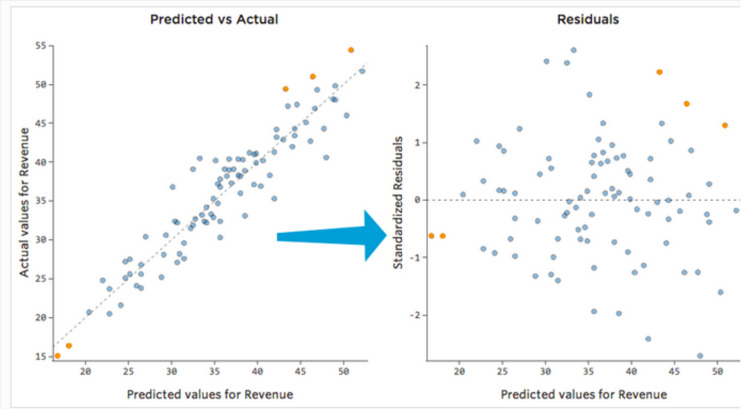
This is an example of a residual plot where there is an outlier



If we have an outlier, it is important to try to identify *why* the observation is an outlier (it could actually be informative). If the point corresponds to a genuine measurement error, then you should remove the outlier, otherwise, see if a transformation (e.g. square-root or log) diminishes the impact of the outlier.

Model Diagnostics

Here we can see how the predicted/fitted versus observed plot is transformed into a residual plot



Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

Exercise

Exercise: Model Diagnostics

For our cigarette example, plot a fitted versus observed plot as well as a residual plot. Do you think the model is fitting well?