Lecture 3: Linear Regression

STATS 202: Data mining and analysis

Rajan Patel

Simple linear regression

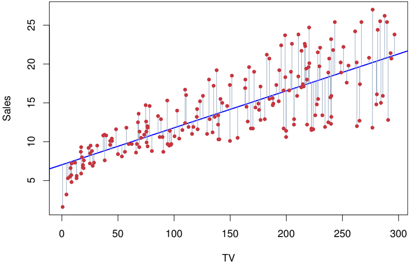


Figure: *

Figure 3.1

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

Simple linear regression

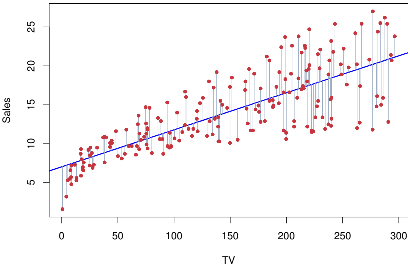


Figure: *

Figure 3.1

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize the residual sum of squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Simple linear regression

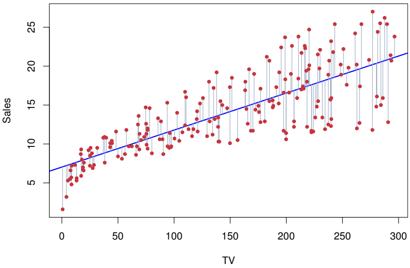


Figure: *

Figure 3.1

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize the residual sum of squares (RSS):

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \end{aligned}$$

Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

A little calculus shows that the minimizers of the RSS are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

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

Assesing the accuracy of $\hat{\beta}_0$ and $\hat{\beta}_1$

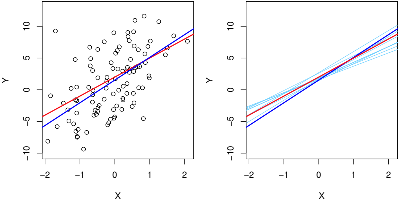


Figure: *

Figure 3.3

The Standard Errors for the parameters are:

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right]$$

$$\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Assesing the accuracy of $\hat{\beta}_0$ and $\hat{\beta}_1$

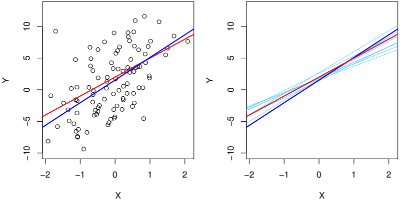


Figure: *

Figure 3.3

The Standard Errors for the parameters are:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

The 95% confidence intervals:

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Hypothesis test

H_0: There is no relationship between X and Y.

H_a: There is some relationship between X and Y.

Hypothesis test

H_0: $\beta_1 = 0$.

H_a: $\beta_1 \neq 0$.

Hypothesis test

H_0: $\beta_1 = 0$.

H_a: $\beta_1 \neq 0$.

Test statistic: $t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)}$.

Under the null hypothesis, this has a t-distribution with $n-2$ degrees of freedom.

Hypothesis test

$$H_0\colon\ \beta_1 = 0.$$

$$H_a\colon\ \beta_1 \neq 0.$$

Test statistic: $\quad t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)}.$

Under the null hypothesis, this has a t-distribution with $n-2$ degrees of freedom.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* `Advertising` *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of \$1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* `sales` *variable is in thousands of units, and the* `TV` *variable is in thousands of dollars).*

Interpreting the hypothesis test

- If we reject the null hypothesis, can we assume there is a linear relationship?

Interpreting the hypothesis test

- If we reject the null hypothesis, can we assume there is a linear relationship?
  - No. A quadratic relationship may be a better fit, for example.

Interpreting the hypothesis test

- If we reject the null hypothesis, can we assume there is a linear relationship?
  - No. A quadratic relationship may be a better fit, for example.
- If we don't reject the null hypothesis, can we assume there is no relationship between X and Y?

Interpreting the hypothesis test

- If we reject the null hypothesis, can we assume there is a linear relationship?
  - No. A quadratic relationship may be a better fit, for example.
- If we don't reject the null hypothesis, can we assume there is no relationship between X and Y?
  - No. This test is only powerful against certain monotone alternatives. There could be more complex non-linear relationships.

Multiple linear regression

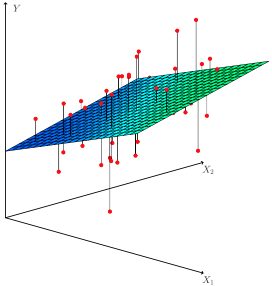


Figure: *

Figure 3.4

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

Multiple linear regression

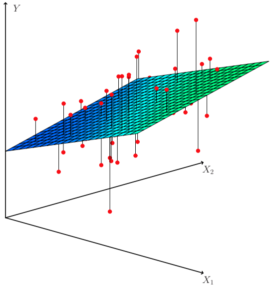


Figure: *

Figure 3.4

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

or, in matrix notation:

$$E\mathbf{y} = \mathbf{X}\beta,$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\beta = (\beta_0, \ldots, \beta_p)^T$ and $\mathbf{X}$ is our usual data matrix with an extra column of ones on the left to account for the intercept.

Multiple linear regression answers several questions

- Is at least one of the variables X_i useful for predicting the outcome Y?

Multiple linear regression answers several questions

- Is at least one of the variables X_i useful for predicting the outcome Y?
- Which subset of the predictors is most important?

Multiple linear regression answers several questions

- Is at least one of the variables X_i useful for predicting the outcome Y?
- Which subset of the predictors is most important?
- How good is a linear model for these data?

Multiple linear regression answers several questions

- Is at least one of the variables X_i useful for predicting the outcome Y?
- Which subset of the predictors is most important?
- How good is a linear model for these data?
- Given a set of predictor values, what is a likely value for Y, and how accurate is this prediction?

The estimates $\hat{\beta}$

Our goal again is to minimize the RSS:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The estimates $\hat{\beta}$

Our goal again is to minimize the RSS:

$$
\begin{aligned}
\mathsf{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2.
\end{aligned}
$$

The estimates $\hat{\beta}$

Our goal again is to minimize the RSS:

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2. \end{aligned}$$

One can show that this is minimized by the vector $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Which variables are important?

Consider the hypothesis:

H_0 : The last q predictors have no relation with Y.

Which variables are important?

Consider the hypothesis:

$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$

Which variables are important?

Consider the hypothesis:

$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$

Let RSS_0 be the residual sum of squares for the model which excludes these variables.

Which variables are important?

Consider the hypothesis:

$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F-statistic is defined by:

$$F = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/q}{\mathsf{RSS}/(n-p-1)}.$$

Which variables are important?

Consider the hypothesis:

$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F-statistic is defined by:

$$F = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/q}{\mathsf{RSS}/(n-p-1)}.$$

Under the null hypothesis, this has an F-distribution.

Which variables are important?

Consider the hypothesis:

$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F-statistic is defined by:

$$F = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/q}{\mathsf{RSS}/(n-p-1)}.$$

Under the null hypothesis, this has an F-distribution.

Example: If $q = p$, we test whether any of the variables is important.

$$\mathsf{RSS}_0 = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Which variables are important?

A multiple linear regression in R has the following output:

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.594  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.761e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-Squared: 0.7406,     Adjusted R-squared: 0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Which variables are important?

The t-statistic associated to the ith predictor is the square root of the F-statistic for the null hypothesis which sets only $\beta_i = 0$.

Which variables are important?

The t-statistic associated to the ith predictor is the square root of the F-statistic for the null hypothesis which sets only $\beta_i = 0$.

A low p-value indicates that the predictor is important.

Which variables are important?

The t-statistic associated to the ith predictor is the square root of the F-statistic for the null hypothesis which sets only $\beta_i = 0$.

A low p-value indicates that the predictor is important.

Warning: If there are many predictors, even under the null hypothesis, some of the t-tests will have low p-values.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

A way to simplify the choice is to define a range of models with an increasing number of variables, then select the best.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

A way to simplify the choice is to define a range of models with an increasing number of variables, then select the best.

- **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

A way to simplify the choice is to define a range of models with an increasing number of variables, then select the best.

- **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest p-value at each step.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

A way to simplify the choice is to define a range of models with an increasing number of variables, then select the best.

- **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest p-value at each step.
- **Mixed selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate that variable.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

A way to simplify the choice is to define a range of models with an increasing number of variables, then select the best.

- **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest p-value at each step.
- **Mixed selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate that variable.

Choosing one model in the range produced is a form of *tuning*.

How good is the fit?

To assess the fit, we focus on the residuals.

How good is the fit?

To assess the fit, we focus on the residuals.

- The RSS always decreases as we add more variables.

How good is the fit?

To assess the fit, we focus on the residuals.

- The RSS always decreases as we add more variables.
- The residual standard error (RSE) corrects this:

$$\mathsf{RSE} = \sqrt{\frac{1}{n-p-1}\mathsf{RSS}}.$$

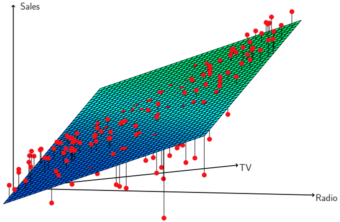
How good is the fit?

To assess the fit, we focus on the residuals.

- The RSS always decreases as we add more variables.
- The residual standard error (RSE) corrects this:

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}.$$

- Visualizing the residuals can reveal phenomena that are not accounted for by the model; eg. synergies or interactions:



How good are the predictions?

The function `predict` in R output predictions from a linear model:

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
        interval="confidence")
    fit   lwr   upr
1 29.80 29.01 30.60
2 25.05 24.47 25.63
3 20.30 19.73 20.87
```

Confidence intervals reflect the uncertainty on $\hat{\beta}$.

How good are the predictions?

The function `predict` in R output predictions from a linear model:

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
       interval="confidence")
    fit    lwr    upr
1 29.80 29.01 30.60
2 25.05 24.47 25.63
3 20.30 19.73 20.87
```

Confidence intervals reflect the uncertainty on $\hat{\beta}$.

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
       interval="prediction")
    fit     lwr   upr
1 29.80 17.566 42.04
2 25.05 12.828 37.28
3 20.30  8.078 32.53
```

Prediction intervals reflect uncertainty on $\hat{\beta}$ and the irreducible error ε as well.

Multiple linear regression

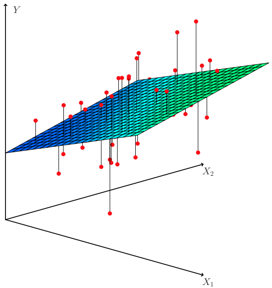


Figure: *

Figure 3.4

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

or, in matrix notation:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\beta = (\beta_0, \ldots, \beta_p)^T$ and $\mathbf{X}$ is our usual data matrix with an extra column of ones on the left to account for the intercept.

The estimates $\hat{\beta}$

Our goal is to minimize the RSS (training error):

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2.\end{aligned}$$

This is minimized by the vector $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

This only exists when $\mathbf{X}^T\mathbf{X}$ is invertible. This requires $n \geq p$.

Testing whether a group of variables is important

- F-test:

  $$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

  RSS_0 is the residual sum of squares for the model in H_0.

  $$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}.$$

- Special case: $q = p$. Test whether any of the predictors are important.
- Special case: $q = 1$, exclude a single variable. Test whether this variable is important $\sim$ t-tests in R output. Must be careful with multiple testing.

How many variables are important?

When choosing a subset of the predictors, we have 2^p choices. We cannot test every possible subset!

Instead we will use a **stepwise approach**:

1. Construct a sequence of p models with increasing number of variables.
2. Select the best model among them.

Three variants of stepwise selection

- **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest t-test p-value at each step.
- **Mixed selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate that variable.

How many variables are important?

The output of a stepwise selection method is a range of models:

- $\{\}$
- $\{\texttt{tv}\}$
- $\{\texttt{tv}, \texttt{newspaper}\}$
- $\{\texttt{tv}, \texttt{newspaper}, \texttt{radio}\}$
- $\{\texttt{tv}, \texttt{newspaper}, \texttt{radio}, \texttt{facebook}\}$
- $\{\texttt{tv}, \texttt{newspaper}, \texttt{radio}, \texttt{facebook}, \texttt{twitter}\}$

6 choices are better than $2^6 = 64$. We use different *tuning methods* to decide which model to use; e.g. cross-validation, AIC, BIC.

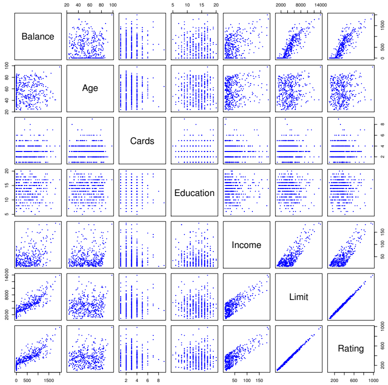
How many variables are important?

When choosing a subset of the predictors, we have 2^p choices.

- **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest p-value at each step.
- **Mixed selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate that variable.

Dealing with categorical or qualitative predictors

Example: Credit dataset



In addition, there are 4 qualitative variables:

- `gender`: male, female.
- `student`: student or not.
- `status`: married, single, divorced.
- `ethnicity`: African American, Asian, Caucasian.

Dealing with categorical or qualitative predictors

For each qualitative predictor, e.g. `ethnicity`:

- Choose a baseline category, e.g. African American
- For every other category, define a new predictor:
  - X_{Asian} is 1 if the person is Asian and 0 otherwise.
  - $X_{\text{Caucasian}}$ is 1 if the person is Caucasian and 0 otherwise.

The model will be:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + \beta_{\text{Asian}} X_{\text{Asian}} + \beta_{\text{Caucasian}} X_{\text{Caucasian}} + \varepsilon.$$

β_{Asian} is the relative effect on `balance` for being Asian compared to the baseline category.

Dealing with categorical or qualitative predictors

- The model fit and predictions are independent of the choice of the baseline category.
- However, hypothesis tests derived from these variables are affected by the choice.
  - **Solution:** To check whether `ethnicity` is important, use an F-test for the hypothesis $\beta_{\text{Asian}} = \beta_{\text{Caucasian}} = 0$. This does not depend on the coding.
- Other ways to encode qualitative predictors produce the same fit $\hat{f}$, but the coefficients have different interpretations.

How good are the predictions?

The function `predict` in R output predictions from a linear model; eg. $x_0 = (5, 10, 15)$:

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
       interval="confidence")
    fit   lwr   upr
1 29.80 29.01 30.60
2 25.05 24.47 25.63
3 20.30 19.73 20.87
```

"Confidence intervals" reflect the uncertainty on $\hat{\beta}$; ie. confidence interval for $f(x_0)$.

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
       interval="prediction")
    fit    lwr   upr
1 29.80 17.566 42.04
2 25.05 12.828 37.28
3 20.30  8.078 32.53
```

"Prediction intervals" reflect uncertainty on $\hat{\beta}$ and the irreducible error ε as well; i.e. confidence interval for y_0.

Recap

So far, we have:

- Defined Multiple Linear Regression
- Discussed how to test the importance of variables.
- Described one approach to choose a subset of variables.
- Explained how to code qualitative variables.
- Now, how do we evaluate model fit? Is the linear model any good? What can go wrong?

How good is the fit?

To assess the fit, we focus on the residuals.

- $R^2 = \text{Corr}(Y, \hat{Y})$, always increases as we add more variables.
- The residual standard error (RSE) does not always improve with more predictors:

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}.$$

- **Visualizing the residuals** can reveal phenomena that are not accounted for by the model.

Potential issues in linear regression

1. Interactions between predictors
2. Non-linear relationships
3. Correlation of error terms
4. Non-constant variance of error (heteroskedasticity).
5. Outliers
6. High leverage points
7. Colinearity

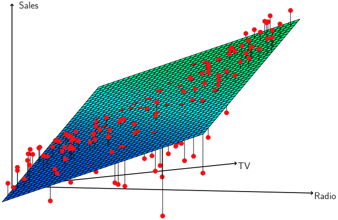
Interactions between predictors

Linear regression has an *additive* assumption:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{tv} + \beta_2 \times \texttt{radio} + \varepsilon$$

i.e. An increase of \$100 dollars in TV ads causes a fixed increase in sales, regardless of how much you spend on radio ads.

If we visualize the residuals, it is clear that this is false:



Interactions between predictors

One way to deal with this is to include multiplicative variables in the model:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{tv} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{tv} \cdot \texttt{radio}) + \varepsilon$$

The **interaction variable** is high when both `tv` and `radio` are high.

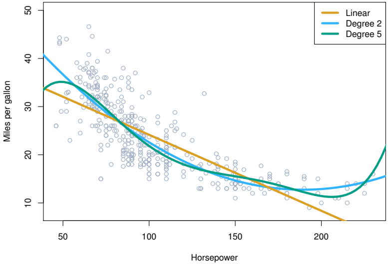
Interactions between predictors

R makes it easy to include interaction variables in the model:

```
> lm.fit=lm(Sales~.+Income:Advertising+Price:Age,data=Carseats)
> summary(lm.fit)

Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data =
     Carseats)

Residuals:
   Min     1Q  Median     3Q    Max
-2.921 -0.750  0.018  0.675  3.341

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.575565   1.008747    6.52  2.2e-10 ***
CompPrice           0.092937   0.004118   22.57  < 2e-16 ***
Income              0.010894   0.002604    4.18  3.6e-05 ***
Advertising         0.070246   0.022609    3.11  0.00203 **
Population          0.000159   0.000368    0.43  0.66533
Price              -0.100806   0.007440  -13.55  < 2e-16 ***
ShelveLocGood       4.848676   0.152838   31.72  < 2e-16 ***
ShelveLocMedium     1.953262   0.125768   15.53  < 2e-16 ***
Age                -0.057947   0.015951   -3.63  0.00032 ***
Education          -0.020852   0.019613   -1.06  0.28836
UrbanYes            0.140160   0.112402    1.25  0.21317
USYes              -0.157557   0.148923   -1.06  0.29073
Income:Advertising  0.000751   0.000278    2.70  0.00729 **
Price:Age           0.000107   0.000133    0.80  0.42381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Non-linearities

Example: Auto dataset.



A scatterplot between a predictor and the response may reveal a non-linear relationship.

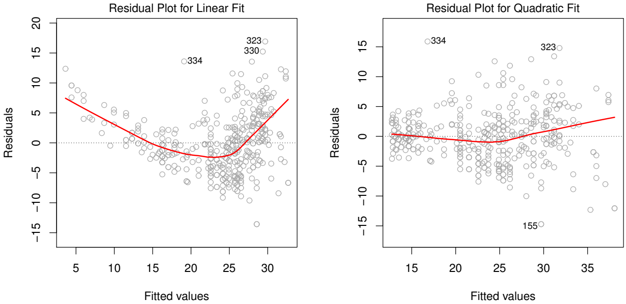
Solution: include polynomial terms in the model.

$$
\begin{aligned}
\texttt{MPG} = \beta_0 &+ \beta_1 \times \texttt{horsepower} + \varepsilon \\
&+ {\color{red}\beta_2 \times \texttt{horsepower}^2} + \varepsilon \\
&+ {\color{red}\beta_3 \times \texttt{horsepower}^3} + \varepsilon \\
&+ \ldots + \varepsilon
\end{aligned}
$$

Non-linearities

In 2 or 3 dimensions, this is easy to visualize. What do we do when we have too many predictors?

Plot the residuals against the *response* and look for a pattern:



Correlation of error terms

We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma) \text{ i.i.d.}$$

What if this breaks down?

The main effect is that this invalidates any assertions about Standard Errors, confidence intervals, and hypothesis tests:

Example: Suppose that by accident, we double the data (we use each sample twice). Then, the standard errors would be artificially smaller by a factor of $\sqrt{2}$.

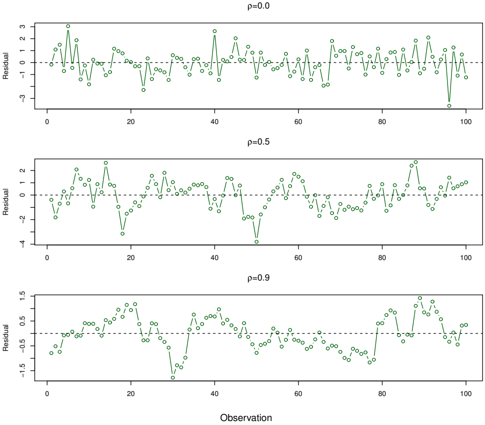
Correlation of error terms

When could this happen in real life:

- **Time series:** Each sample corresponds to a different point in time. The errors for samples that are close in time are correlated.
- **Spatial data:** Each sample corresponds to a different location in space.
- Study on predicting height from weight at birth. Suppose some of the subjects in the study are in the same family, their shared environment could make them deviate from $f(x)$ in similar ways.

Correlation of error terms

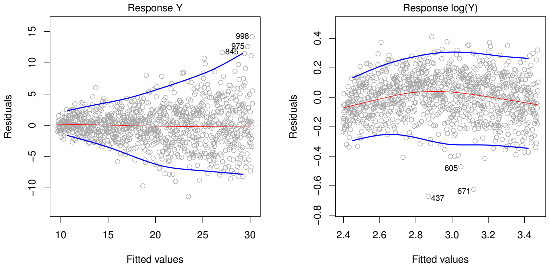
Simulations of time series with increasing correlations between ε_i.



Non-constant variance of error (heteroskedasticity)

The variance of the error depends on the input.

To diagnose this, we can plot residuals vs. fitted values:



Solution: If the trend in variance is relatively simple, we can transform the response using a logarithm, for example.