Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
Outline

Introduction

Simple Linear Regression

Multiple Linear Regression

Other Considerations in the Regression Model
Simple and widely used approach for regression (i.e. supervised learning with quantitative response)

More sophisticated methods are “generalizations” of linear regression

Goals of this chapter:

- Review key ideas underlying the linear regression model
- Study the least squares approach
Example: Advertising data set

Figure 1: Sales (in thousands of units) as a function of TV, radio, and newspaper budgets (in thousands of dollars), for 200 markets. In each plot, the blue line represents the least squares fit of sales to that variable.
Questions of interest

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?
Outline

Introduction

Simple Linear Regression

Multiple Linear Regression

Other Considerations in the Regression Model
Simple Linear Regression

- Simple linear regression assumes a linear relationship between a single predictor $X$ and a quantitative response $Y$

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

where $\varepsilon$ is random error term, which is independent of $X$ and has mean 0

- $\beta_0$ (intercept), $\beta_1$ (slope) are the model coefficients or parameters

- Example: $X$ is TV and $Y$ is sales

\[ \text{sales} = \beta_0 + \beta_1 \times \text{TV} + \varepsilon \]

- Use training data to get estimates $\hat{\beta}_0$, $\hat{\beta}_1$. Prediction for $X = x$ is then given by

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]
Estimating the Coefficients

- Training data: \((x_1, y_1), \ldots, (x_n, y_n)\)
- In **Advertising** data set, \(n = 200\) different markets
- **Goal:** Find \(\hat{\beta}_0, \hat{\beta}_1\) such that \(y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i\), for all \(i = 1, \ldots, n\)
- 2 methods of estimation:
  1. Minimizing the least squares criterion
  2. Maximizing the likelihood
Least Squares

- Prediction: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- The $i$-th residual: $e_i = y_i - \hat{y}_i$
- Residual Sum of Squares: $\text{RSS} = \sum_{i=1}^{n} e_i^2$, or equivalently
  \[ \text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \]
- Least squares approach chooses $\hat{\beta}_0, \hat{\beta}_1$ that minimize RSS

Property (Least Squares Coefficient Estimates)

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]
\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ are the sample means
Definition (Likelihood function)

Let \( X_1, \ldots, X_n \) be a random sample from a distribution that depends on one or more unknown parameters \( \theta_1, \ldots, \theta_m \) with probability density (or mass) function \( p(x_i; \theta_1, \ldots, \theta_m) \). Suppose \( (\theta_1, \ldots, \theta_m) \) is restricted to a parameter space \( \Theta \subseteq \mathbb{R}^m \). Then, when regarded as a function of \( \theta_1, \ldots, \theta_m \), the joint probability density (or mass) function of \( X_1, \ldots, X_n \):

\[
L(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{m} p(x_i; \theta_1, \ldots, \theta_m)
\]

is called the likelihood function.
Maximum Likelihood Estimation

Definition (Likelihood function)

Let $X_1, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \ldots, \theta_m$ with probability density (or mass) function $p(x_i; \theta_1, \ldots, \theta_m)$.

1. If:

\[
(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n))
\]

is an $m$-tuple that maximizes the likelihood function $L(\theta_1, \ldots, \theta_m)$, then:

\[
\hat{\theta}_j = g_j(X_1, \ldots, X_n)
\]

is a maximum likelihood estimator (MLE) of $\theta_j$ for $j = 1, \ldots, m$

2. The corresponding observed values of the statistics in (1), namely: $(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n))$ are called the maximum likelihood estimates of $\theta_j$ for $j = 1, \ldots, m$
Maximum Likelihood Estimation

Methodology: Given a random sample $X_1, \ldots, X_n$ with probability density (or mass) function $p(x_i; \theta_1, \ldots, \theta_m)$

1. Compute the likelihood function $L(\theta_1, \ldots, \theta_m)$
2. Compute $\ln L(\theta_1, \ldots, \theta_m)$ the logarithm of the likelihood function
3. To find $\hat{\theta}_j$ an MLE for $\theta_j$, $j = 1, \ldots, m$, take the partial derivative of $\ln L(\theta_1, \ldots, \theta_m)$ with respect to $\theta_j$, and set to 0
Interpreting the Coefficient Estimates

Figure 2: Simple Linear Regression of sales onto TV.\
\[ \hat{\beta}_0 \approx 7.03, \hat{\beta}_1 \approx 0.0475. \]

- Recall model: \( Y = \beta_0 + \beta_1 X + \epsilon \), with \( \epsilon \) mean-zero random error term
- \( \beta_0 = \mathbb{E}[Y|X = 0] \)
- \( \beta_1 \) is the average increase of \( Y \) with a 1-unit increase of \( X \)
Figure 3: Generation of 100 $Y$s and $X$s from the model $Y = 2 + 3X + \varepsilon$. Red line is the true population regression line $f(X) = 2 + 3X$. Blue lines are least squares lines computed from different random sets of observations.
Assessing the Accuracy of the Coefficient Estimates

- The true relationship, $f(X) = \beta_0 + \beta_1 X$, is generally not known for real data.
- We can only use information from a sample to estimate characteristics of a population.
- $\hat{\beta}_0$, $\hat{\beta}_1$ are empirical estimates of unknown coefficients $\beta_0$, $\beta_1$.

Property

The least square coefficient estimates are unbiased, that is

$$\mathbb{E}[\hat{\beta}_0] = \beta_0$$
$$\mathbb{E}[\hat{\beta}_1] = \beta_1$$

- How close are $\hat{\beta}_0$ and $\hat{\beta}_1$ to $\beta_0$ and $\beta_1$?
Assessing the Accuracy of the Coefficient Estimates

Property (Standard Errors (SE) of $\hat{\beta}_0$ and $\hat{\beta}_1$)

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

\[
\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

where $\sigma^2 = \text{Var}(\varepsilon)$.

▶ In general, $\sigma^2$ is unknown. An estimate for $\sigma$ is known as the residual standard error: \( \text{RSE} = \sqrt{\text{RSS}/(n-2)} \).
Assessing the Accuracy of the Coefficient Estimates

Property (95% Confidence intervals for $\beta_0$ and $\beta_1$)

\[
P\left(\beta_1 \in \left[ \hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1) \right]\right) \approx 0.95
\]
\[
P\left(\beta_0 \in \left[ \hat{\beta}_0 - 2 \cdot \text{SE}(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot \text{SE}(\hat{\beta}_0) \right]\right) \approx 0.95
\]

- In the case of **Advertising** data, 95% confidence intervals for $\beta_0$ and $\beta_1$ are $[6.130, 7.935]$ and $[0.042, 0.053]$, respectively
Assessing the Accuracy of the Coefficient Estimates

- Perform the **hypothesis test**:

  \[ H_0 : \beta_1 = 0 \text{ (no relationship between } X \text{ and } Y) \]
  \[ H_a : \beta_1 \neq 0 \text{ (there is some relationship between } X \text{ and } Y) \]

- How to perform the hypothesis test?
  - The \( t \)-statistic: \( T = \hat{\beta}_1 / \text{SE}(\hat{\beta}_1) \) follows a \( t \)-distribution with \( n - 2 \) degrees of freedom, assuming \( H_0 \)
  - Compute the \( t \)-statistic \( t_n \) on training data
  - We call **p-value** the probability \( \mathbb{P}(T \geq |t_n| \mid H_0) \)
  - Reject \( H_0 \) if the p-value is small enough (typical cutoffs are 5% or 1%)

- **Advertising** data: Simple linear regression of **sales** onto **TV**

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>SE</th>
<th>( t_n )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.0325</td>
<td>0.4578</td>
<td>15.36</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>TV</strong></td>
<td>0.0475</td>
<td>0.0027</td>
<td>17.67</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Assessing the Accuracy of the Model

The quality of a linear regression fit is assessed using two quantities

1. The residual standard error:

\[
RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}
\]

2. The \(R^2\) statistic:

\[
R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}
\]

where \(TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2\) is the total sum of squares

Recall the empirical correlation coefficient:

\[
\text{corr}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}
\]

In the simple linear regression setting, \(R^2 = \text{Cor}(X, Y)^2\)
Outline

Introduction

Simple Linear Regression

Multiple Linear Regression

Other Considerations in the Regression Model
Multiple Linear Regression

- We want to extend the simple linear regression when we have \( p \) predictors \( X_1, \ldots, X_p \), with \( p \geq 2 \)
- Example: predict sales based on money spent advertising on TV, radio, newspaper
- What about running several separate simple linear regressions?
  - Prediction?
  - Correlations among predictors?
- The multiple linear regression model with \( p \) predictors
  \[
  Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon
  \]
- For the Advertising data,
  \[
  sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \varepsilon
  \]
Estimating the Coefficients

- Model: \( Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \)
- Obtain estimates \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \) of \( \beta_0, \beta_1, \ldots, \beta_p \) using training data
- Prediction: \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p \)
- Least squares approach: choose \( \beta_0, \beta_1, \ldots, \beta_p \) that minimize \( \text{RSS} \), where

\[
\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2
\]
Figure 4: Synthetic data with two predictors and one response. The plane corresponds to the least squares fit.
Matrix form

- The model can be rewritten as: \( y = X\beta + \varepsilon \) where

\[
\begin{pmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{pmatrix}
\in \mathbb{R}^n \text{ is the vector of observations}
\]

\[
X = 
\begin{pmatrix}
  1 & x_{11} & x_{12} & \cdots & x_{1p} \\
  1 & x_{21} & x_{22} & \cdots & x_{2p} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  1 & x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\in \mathbb{R}^{n \times (p+1)} \text{ is the design matrix}
\]

\[
\beta = 
\begin{pmatrix}
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_p
\end{pmatrix}
\in \mathbb{R}^{p+1} \text{ is the regression vector}
\]

\[
\varepsilon = 
\begin{pmatrix}
  \varepsilon_1 \\
  \vdots \\
  \varepsilon_n
\end{pmatrix}
\in \mathbb{R}^n \text{ is the noise vector}
\]
Least Squares Coefficients

- Model (matrix form): \( y = X\beta + \varepsilon \)

**Theorem**

The least squares coefficient estimates are given by

\[
\hat{\beta} = (X^T X)^{-1} X^T y
\]

where \( \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \in \mathbb{R}^{p+1} \)
Simple vs Multiple Regression

- **Simple** linear regression of **sales** onto **TV**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SE</th>
<th>$t_n$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.0325</td>
<td>0.4578</td>
<td>15.36</td>
</tr>
<tr>
<td>TV</td>
<td>0.0475</td>
<td>0.0027</td>
<td>17.67</td>
</tr>
</tbody>
</table>

- **Simple** linear regression of **sales** onto **radio**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SE</th>
<th>$t_n$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.312</td>
<td>0.563</td>
<td>16.54</td>
</tr>
<tr>
<td>radio</td>
<td>0.203</td>
<td>0.020</td>
<td>9.92</td>
</tr>
</tbody>
</table>

- **Simple** linear regression of **sales** onto **newspaper**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SE</th>
<th>$t_n$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>12.351</td>
<td>0.621</td>
<td>19.88</td>
</tr>
<tr>
<td>newspaper</td>
<td>0.055</td>
<td>0.017</td>
<td>3.30</td>
</tr>
</tbody>
</table>

- **Multiple** linear regression of **sales** onto **TV**, **radio**, and **newspaper**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SE</th>
<th>$t_n$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.939</td>
<td>0.3119</td>
<td>9.42</td>
</tr>
<tr>
<td>TV</td>
<td>0.046</td>
<td>0.0014</td>
<td>32.81</td>
</tr>
<tr>
<td>radio</td>
<td>0.189</td>
<td>0.0086</td>
<td>21.89</td>
</tr>
<tr>
<td>newspaper</td>
<td>$-0.001$</td>
<td>0.0059</td>
<td>$-0.18$</td>
</tr>
</tbody>
</table>
Simple vs Multiple Regression

- Why can simple and multiple regression coefficients be quite different?
- Correlation matrix for TV, radio, newspaper, and sales

<table>
<thead>
<tr>
<th></th>
<th>TV</th>
<th>radio</th>
<th>newspaper</th>
<th>sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>1</td>
<td>0.0548</td>
<td>0.0567</td>
<td>0.7822</td>
</tr>
<tr>
<td>radio</td>
<td>1</td>
<td>0.3541</td>
<td>0.5762</td>
<td></td>
</tr>
<tr>
<td>newspaper</td>
<td>1</td>
<td></td>
<td>0.2283</td>
<td></td>
</tr>
<tr>
<td>sales</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Standard Questions

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
Q1: Is There a Relationship Between the Response and Predictors?

- Perform the hypothesis test
  
  $H_0 : \beta_1 = \beta_2 = \ldots = 0$
  
  $H_a :$ at least one $\beta_j$ is non-zero.

- The $F$-statistic
  
  $$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

  follows a $F$-distribution with $(p, n - p - 1)$ degrees of freedom, assuming $H_0$ and if $\varepsilon_i$ are normally distributed

- Compute the $F$-statistic $F_n$ on training data
- Compute the **p-value** $\mathbb{P}(F \geq |F_n| \mid H_0)$
- Reject $H_0$ if the p-value is small enough (typical cutoffs are 5% or 1%)

- **Advertising** data: Multiple linear regression of sales onto TV, newspaper, radio: $F_n = 570$, p-value < 0.0001
Q1: Is There a Relationship Between the Response and Predictors?

- Why an $F$-test instead of $p$ individual $t$-tests?
- Difficulties in the high-dimensional setting: $p > n$
Q2: Deciding on Important Variables

- If $F$-test tells that at least one predictor is associated with the response, which ones?
- **Variable selection**: determining which predictors are associated with the response
- Two challenges:
  1. There are $2^p$ models that contain subsets of $p$ variables
  2. What is the *best* model among all models considered?
- This problem will be studied in a later chapter
Q3: Model Fit ($R^2$)

- Two common measures of fit: $R^2$ and RSE
- $R^2 = \text{Cor}(Y, \hat{Y})^2$. The least-squares method maximizes this correlation among all possible linear models
- If $R^2$ close to 1, the model explains a large portion of the variance in $Y$
- $R^2$ always increases as more variables are added
- Tiny increase in $R^2$ means that a predictor can be dropped from the model

**Advertising** data: 3 models for *sales* using different subsets of predictors

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>0.61</td>
</tr>
<tr>
<td>TV, radio</td>
<td>0.89719</td>
</tr>
<tr>
<td>TV, radio, newspaper</td>
<td>0.8972</td>
</tr>
</tbody>
</table>
Q3: Model Fit (RSE)

- RSE is defined as

\[
RSE = \sqrt{\frac{1}{n - p - 1}} \text{RSS},
\]

(2)

- **Advertising** data: 3 models for sales using different subsets of predictors

<table>
<thead>
<tr>
<th>Model</th>
<th>RSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>3.26</td>
</tr>
<tr>
<td>TV, radio</td>
<td>1.681</td>
</tr>
<tr>
<td>TV, radio, newspaper</td>
<td>1.686</td>
</tr>
</tbody>
</table>

- Note that models with more variables can have higher RSE
Q4: Predictions

- Three sorts of uncertainty associated with the prediction
  1. Assume \( f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \) and
     \( \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p \)
     Inaccuracy in coefficient estimates.
     Compute a confidence interval in order to determine how close
     \( \hat{Y} \) will be to \( f(X) \)
  2. Model bias due to the linear assumption on \( f \)
  3. Random error \( \varepsilon \)

- **Advertising data:** Given that \( TV = \$100,000 \) and \( radio = \$20,000 \)
  - A 95% confidence interval for \( f(X) \) is [10,985, 11,528]
  - A 95% prediction interval for \( Y \) is [7,930, 14,580]

- Prediction intervals are wider than confidence intervals
Qualitative Predictors

- Linear regression model can include *qualitative* independent variables (also known as *factors*).
- Example: predict average credit card debt based on gender, job category.
- We consider two cases depending on the number of possible values (also known as *levels*) for the factors:
  1. Qualitative predictors with only 2 levels
  2. Qualitative predictors with more than 2 levels
Factors With Only 2 Levels

- Create a an indicator or dummy variable that takes on two possible numerical values.
- Example: based on the gender variable, we can create a new variable
  \[ x_i = \begin{cases} 
  1 & \text{if } i \text{-th person is female} \\ 
  0 & \text{if } i \text{-th person is male} 
  \end{cases} \]

- The model becomes:
  \[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} 
  \beta_0 + \beta_1 + \varepsilon_i & \text{if } i \text{-th person is female} \\ 
  \beta_0 + \varepsilon_i & \text{if } i \text{-th person is male} 
  \end{cases} \]

- \( \beta_0 \) is the average credit card balance among males
- \( \beta_0 + \beta_1 \) is the average credit card balance among females
- What if we code females as 0 and males as 1 in (3)?
Factors With More Than 2 Levels

- If a factor has \( L \) levels, create \( L - 1 \) dummy variables
- For instance, **highest degree** has 4 levels: doctorates, masters, bachelors, no college degrees

\[
\begin{align*}
x_{i1} &= \begin{cases} 
1 & \text{if } i\text{-th person’s highest degree is a doctorate} \\
0 & \text{if } i\text{-th person’s highest degree is not a doctorate}
\end{cases} \\
x_{i2} &= \begin{cases} 
1 & \text{if } i\text{-th person’s highest degree is a masters} \\
0 & \text{if } i\text{-th person’s highest degree is not a masters}
\end{cases} \\
x_{i3} &= \begin{cases} 
1 & \text{if } i\text{-th person’s highest degree is a bachelors} \\
0 & \text{if } i\text{-th person’s highest degree is not a bachelors}
\end{cases}
\]

- The level with no dummy variables is known as the baseline
- Choice of the coding has no effect on the regression fit, but only alters the interpretation of the coefficients
Challenging the Additive Assumption

- Additive assumption: the effect of changes in a predictor $X_j$ on the response $Y$ is independent of the values of the other predictors.

- For example, in the case of Advertising data,

  \[ sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \varepsilon \]

- Suppose there is synergy effect: spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.

- This is referred to as interaction effect in statistics.
Relaxing the additive assumption results in the model:

\[
\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{TV} \times \text{radio}) + \varepsilon
\]

\[
= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon
\]

\(\beta_3\) is the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa).

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>SE</th>
<th>(t_n)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.7502</td>
<td>0.248</td>
<td>27.23</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TV</td>
<td>0.0191</td>
<td>0.002</td>
<td>12.70</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>radio</td>
<td>0.0289</td>
<td>0.009</td>
<td>3.24</td>
<td>0.0014</td>
</tr>
<tr>
<td>TV \times radio</td>
<td>0.0011</td>
<td>0.000</td>
<td>20.73</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Hierarchical principle: if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
Challenging the Linear Assumption

- In some cases, the true relationship between the response and the predictors may be **non-linear**
- Simple extension of the linear model is **polynomial regression**

![Figure 5](image_url)

**Figure 5:** horsepower as a function of mpg (gas mileage in miles per gallon). Linear regression fit: regular (orange), including horsepower^2 (blue), including all polynomials of horsepower up to fifth-degree (green).
Incorporating \textit{horsepower}^2 gives

\[ \text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon \] (4)

Model (4) is still a linear model with \( X_1 = \text{horsepower} \) and \( X_2 = \text{horsepower}^2 \)

\begin{center}
\begin{tabular}{l|cccc}
  & Coefficient & SE & \( t_n \) & p-value \\
\hline
Intercept & 56.9001 & 1.8004 & 31.6 & <0.0001 \\
\textit{horsepower} & -0.4662 & 0.0311 & -15.0 & <0.0001 \\
\textit{horsepower}^2 & 0.0012 & 0.0001 & 10.1 & <0.0001 \\
\end{tabular}
\end{center}

\( R^2 = 0.606 \) for the linear fit / \( R^2 = 0.688 \) for the quadratic fit
Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

- Non-linearity of the response-predictor relationships
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity