Linear Regression

IE4213/TIE4213/EE4802
Department of Industrial Systems Engineering & Management

Source: Chapter 3 in JWHT13[1]

January 8, 2021

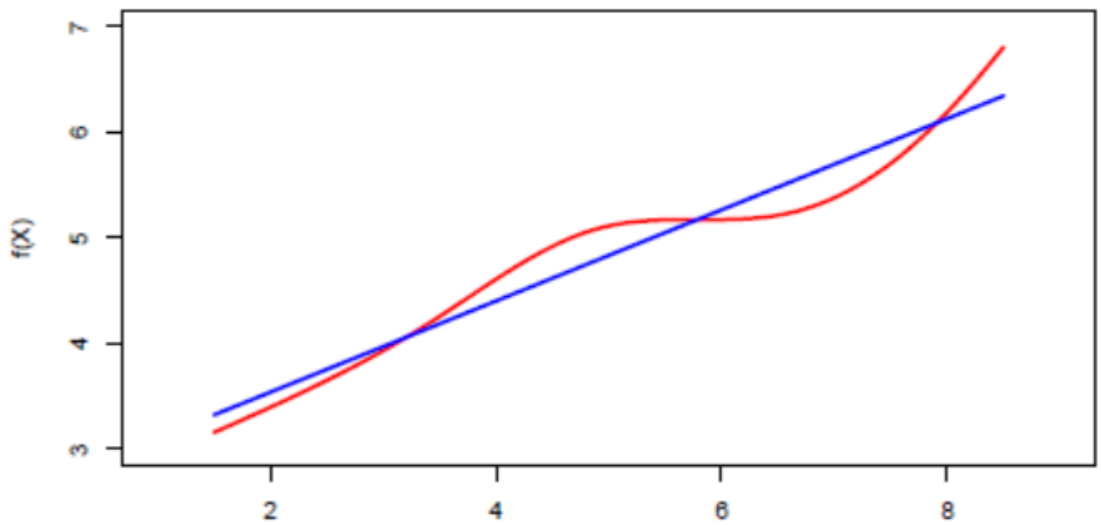
[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer.

Linear Regression: Importance

- Linear regression is a simple approach to supervised learning. It is a useful tool for predicting a quantitative response.
- Though it may seem somewhat dull compared to some of the more modern statistical learning approaches introduced later on, linear regression is still a useful and widely used statistical learning method.
- Moreover, it serves as a good jumping-off point for newer approaches: many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression.
- Consequently, the importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated.
- Objective: review some of the key ideas underlying the linear regression model, and the least squares approach that is most commonly used to fit this model.

Linear Regression

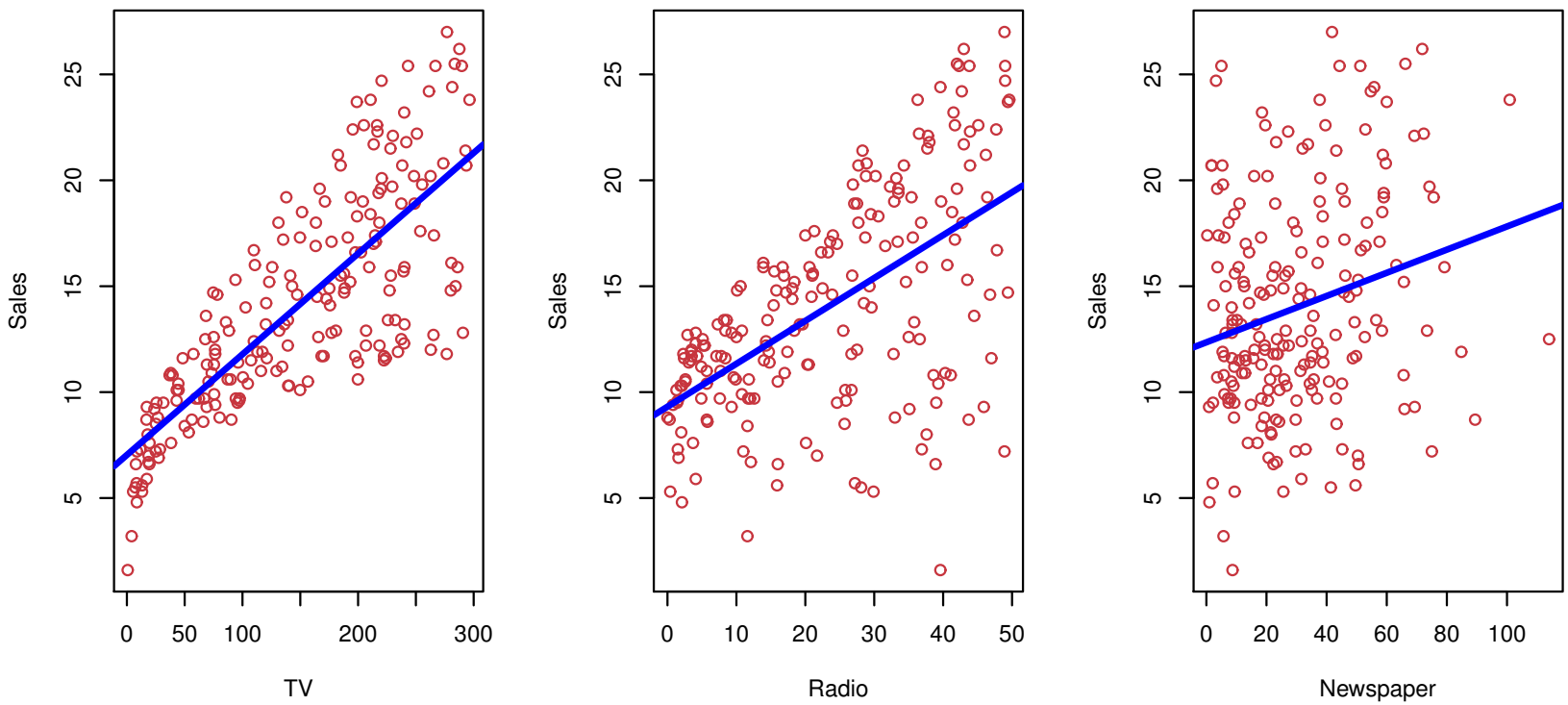
- Linear regression assumes that the dependence of Y on $X_1, X_2, \ldots, X_p$ is linear.
- True regression functions are never linear!



- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Advertising Data

- The figure displays sales (in thousands of units) for a particular product as a function of advertising budgets (in thousands of dollars) for TV , radio , and newspaper media.
- Suppose that in our role as statistical consultants we are asked to suggest, on the basis of this data, a marketing plan for next year that will result in high product sales.



Some Specific Questions I

- Is there a relationship between advertising budget and sales?
  - If the evidence of association is weak, then one might argue that no money should be spent on advertising. (testing model significance)
- How strong is the relation between advertising budget and sales?
  - Given a certain advertising budget, can we predict sales with a high level of accuracy? This would be a strong relationship. Or is a prediction of sales based on advertising expenditure only slightly better than a random guess? This would be a weak relationship.
- Which media contribute to sales?
  - We must find a way to separate out the individual effects of each medium when we have spent money on all three media. (testing covariate significance)
- How accurately can we estimate the effect of a medium on sales?
  - For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase? (uncertainty quantification: confidence interval)

Some Specific Questions II

- How accurately can we predict future sales?
  - For any given level of television, radio, or newspaper advertising, what is our prediction for sales, and what is the accuracy of this prediction? (uncertainty quantification: prediction interval)
- Is the relationship linear?
  - If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still be possible to transform the predictor or the response so that linear regression can be used. (goodness-of-fit test)
- Is there synergy among the advertising media?
  - Perhaps spending \$50,000 on television advertising and \$50,000 on radio advertising results in more sales than allocating \$100,000 to either television or radio individually. In marketing, this is known as a synergy effect, while in statistics it is called an interaction effect.

Overview

Simple Linear Regression

Multiple Linear Regression

Other Considerations in Regression: Qualitative Predictors

Simple LR: Single Predictor X

- Consider a LR model with a single predictor X:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

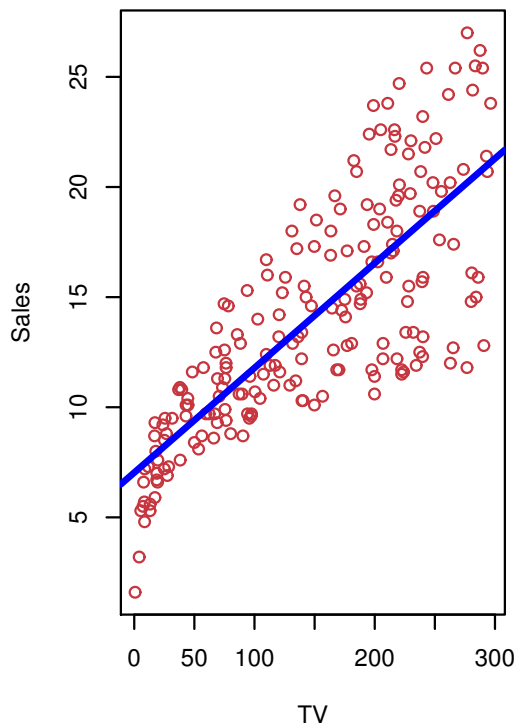
  where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term with mean 0 and (assumed) independent of X.
- The error term is a catch-all for what we miss with this simple model: the influence of all unobserved/unaccounted factors on the response Y. For instance, if we use sales $= \beta_0 + \beta_1 \times$ TV $+ \epsilon$, then ϵ represents the unaccounted impact of radio and newspaper, as well as other unobserved factors.
- We will sometimes describe the above model by saying that we are regressing Y on X (or Y onto X).
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

  where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value.

Estimating the Parameters: Objective

- Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be n observation pairs, each of which consists of a measurement of X and a measurement of Y.
- In the Advertising example, if we use sales $= \beta_0 + \beta_1 \times$ TV $+ \epsilon$, this data set consists of the TV advertising budget and product sales in $n = 200$ different markets.
- Our goal is to obtain coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_1$ s.t. the resulting line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is as close as possible to the $n = 200$ data points: $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \cdots, n$.



Estimating the Parameters: Least Squares

- Measuring the closeness between y_i and $\hat{\beta}_0 + \hat{\beta}_1 x_i$: how?
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual.
- We define the residual sum of squares (RSS) as

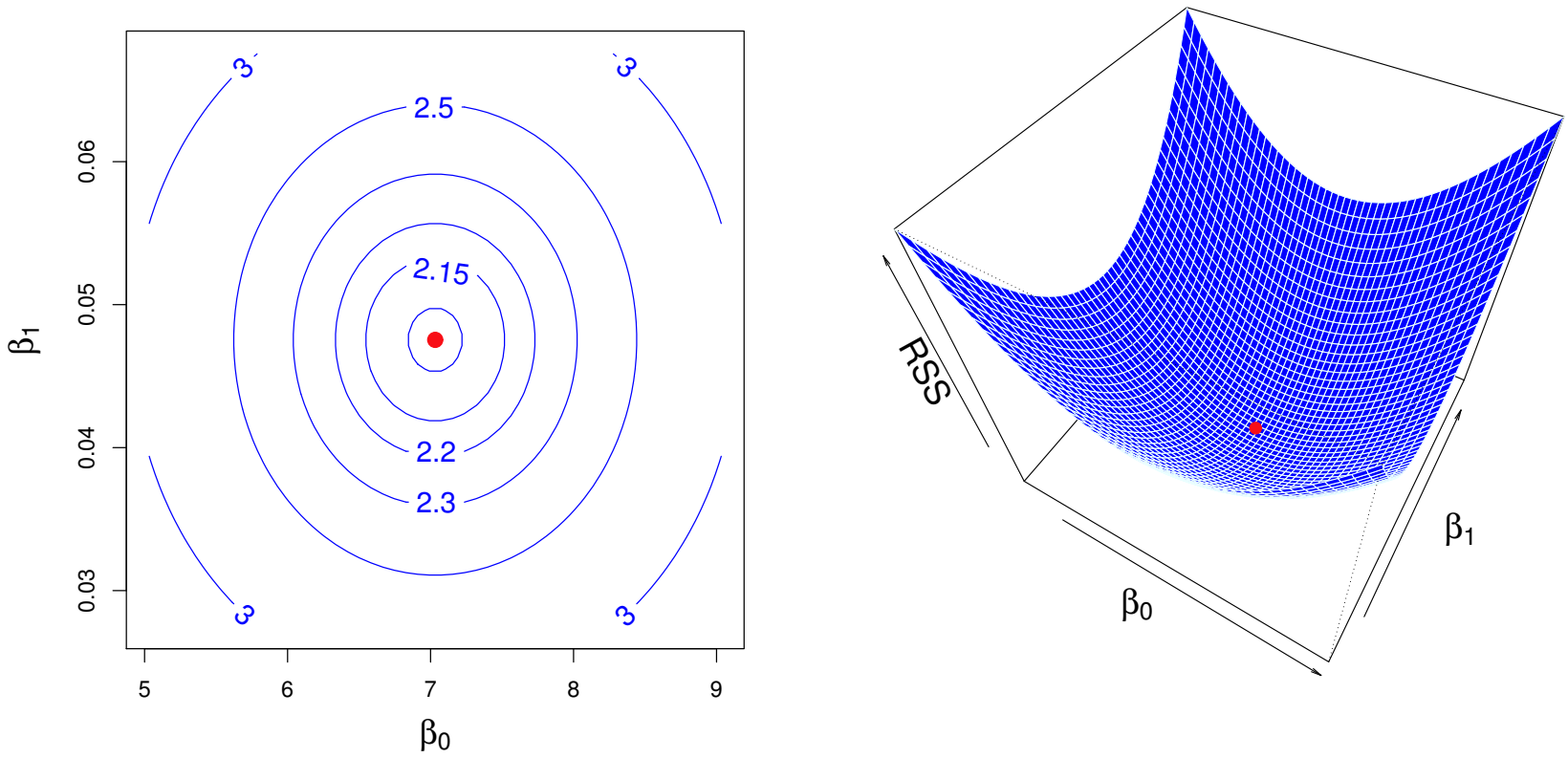
$$\begin{aligned}\texttt{RSS} =& e_1^2 + e_2^2 + \ldots + e_n^2 \\ \texttt{RSS} =& (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2,\end{aligned}$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values (also called least squares coefficient estimates) can be shown to be

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x},\end{aligned}$$

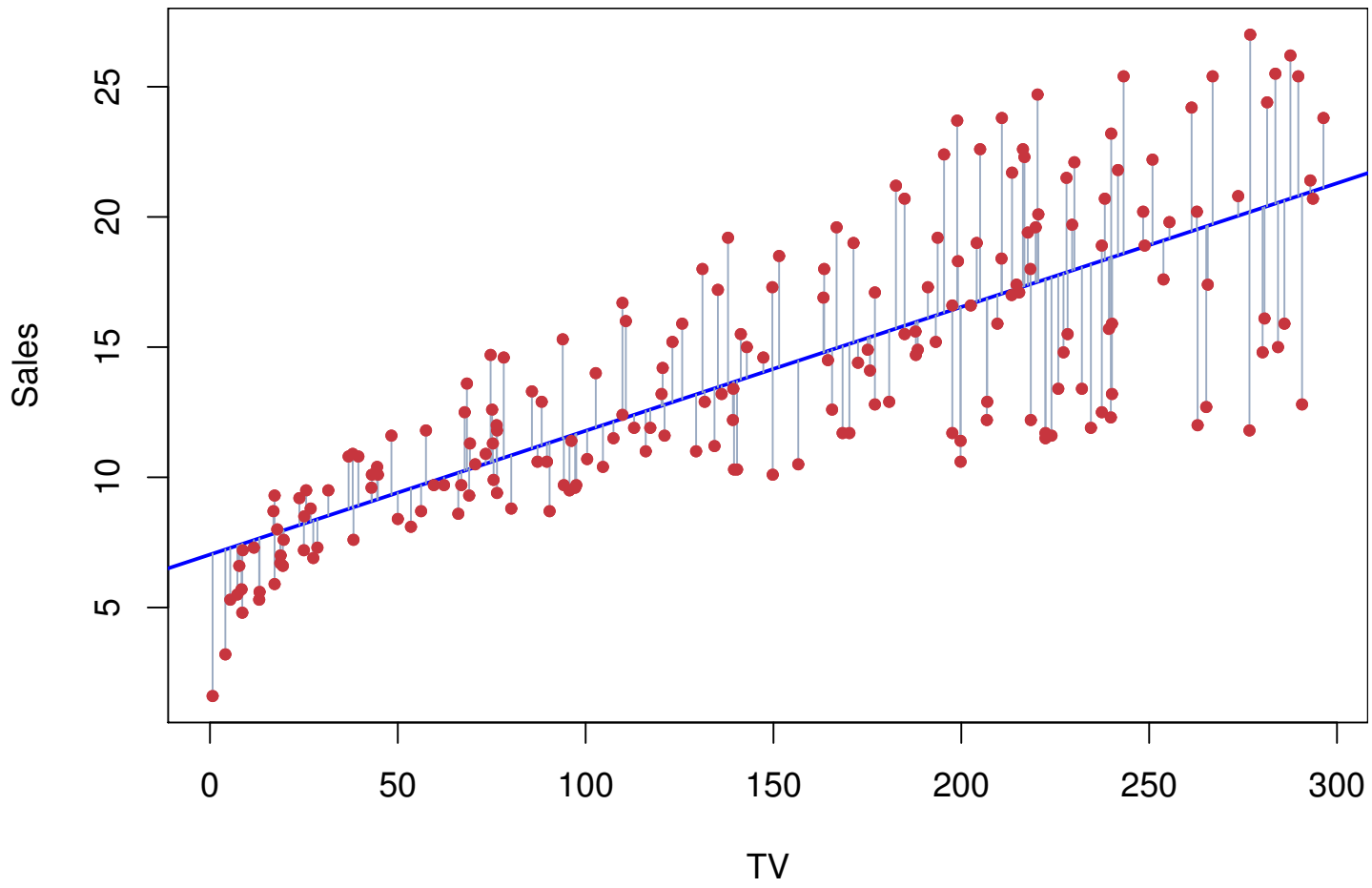
where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ are the sample means.

Plots of the Objective Function



Contour and 3D plots of the RSS on the Advertising data, using `sales` as the response and `TV` as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$: An additional \$1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.

Example: Advertising Data



The least squares fit for the regression of `sales` onto `TV`.
In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Accuracy of the Estimated Coefficient: SE

- Recall: $Y = \beta_0 + \beta_1 X + \epsilon$, where the error term ϵ has mean 0 and is assumed independent of X. Further let $\sigma^2 = \mathrm{Var}(\epsilon)$.
- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

- In general, σ^2 is not known, but can be estimated from the data. The estimate of σ is known as the residual standard error (RSE), and is given by $\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$.
- Strictly speaking, when σ^2 is estimated from the data we should write $\widehat{\mathrm{SE}}(\hat{\beta}_1)^2$ to indicate that an estimate has been made, but for simplicity of notation we will drop this extra hat on SE.

Accuracy of the Estimated Coefficient: CI

- These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form $\hat{\beta}_1 \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}_1)$.
- That is, there is approximately a 95% chance that the interval

$$[\hat{\beta}_1 - 1.96 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \cdot \mathrm{SE}(\hat{\beta}_1)]$$

  will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample).
- In the case of the advertising data, the 95% confidence interval for β_1 is [0.042,0.053].
- Similarly, a confidence interval for β_0 approximately takes the form

$$[\hat{\beta}_0 - 1.96 \cdot \mathrm{SE}(\hat{\beta}_0), \hat{\beta}_0 + 1.96 \cdot \mathrm{SE}(\hat{\beta}_0)].$$

- Advertising data: 95% confidence interval for β_0 is $[6.130, 7.935]$.

Hypothesis Testing I

- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common hypothesis test involves testing the following null hypothesis versus the alternative hypothesis:

$$\text{Null } H_0 : \text{There is no relationship between } X \text{ and } Y,$$
$$\text{Alternative } H_A : \text{There is some relationship between } X \text{ and } Y.$$

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

- To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate for β_1, is sufficiently far from zero that we can be confident that β_1 is non-zero. How far is far enough? This of course depends on the accuracy of $\hat{\beta}_1$

Hypothesis Testing II

- To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.

Results for the advertising data

	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

Assessing the Overall Accuracy

- After we are sure X has some prediction power on Y, the next is to quantify the extent to which the model fits the data.
- The quality of a LR fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.
- Recall: RSS is $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. RSE is defined as

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

- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

  where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.
- It can be shown that in this simple LR setting, $R^2 = r^2$, where r is the correlation coefficient between X and Y:

$$r^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Assessing the Overall Accuracy

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

What is a Good R^2 Value

- The R^2 statistic has an interpretational advantage over the RSE, since unlike the RSE, it always lies between 0 and 1.
- However, it can still be challenging to determine what is a good R^2 value, and in general, this will depend on the application.
- In certain problems in physics, we may know that the data truly comes from a linear model with a small residual error. In this case, we would expect to see an R^2 value that is extremely close to 1, and a substantially smaller R^2 value might indicate a serious problem with the experiment in which the data were generated.
- In typical applications in biology, psychology, marketing, and other domains, the linear model is at best an extremely rough approximation to the data, and residual errors due to other unmeasured factors are often very large. In this setting, we would expect only a very small proportion of the variance in the response to be explained by the predictor, and an R^2 value well below 0.1 might be more realistic!

Overview

Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

Estimation: Least Squares

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

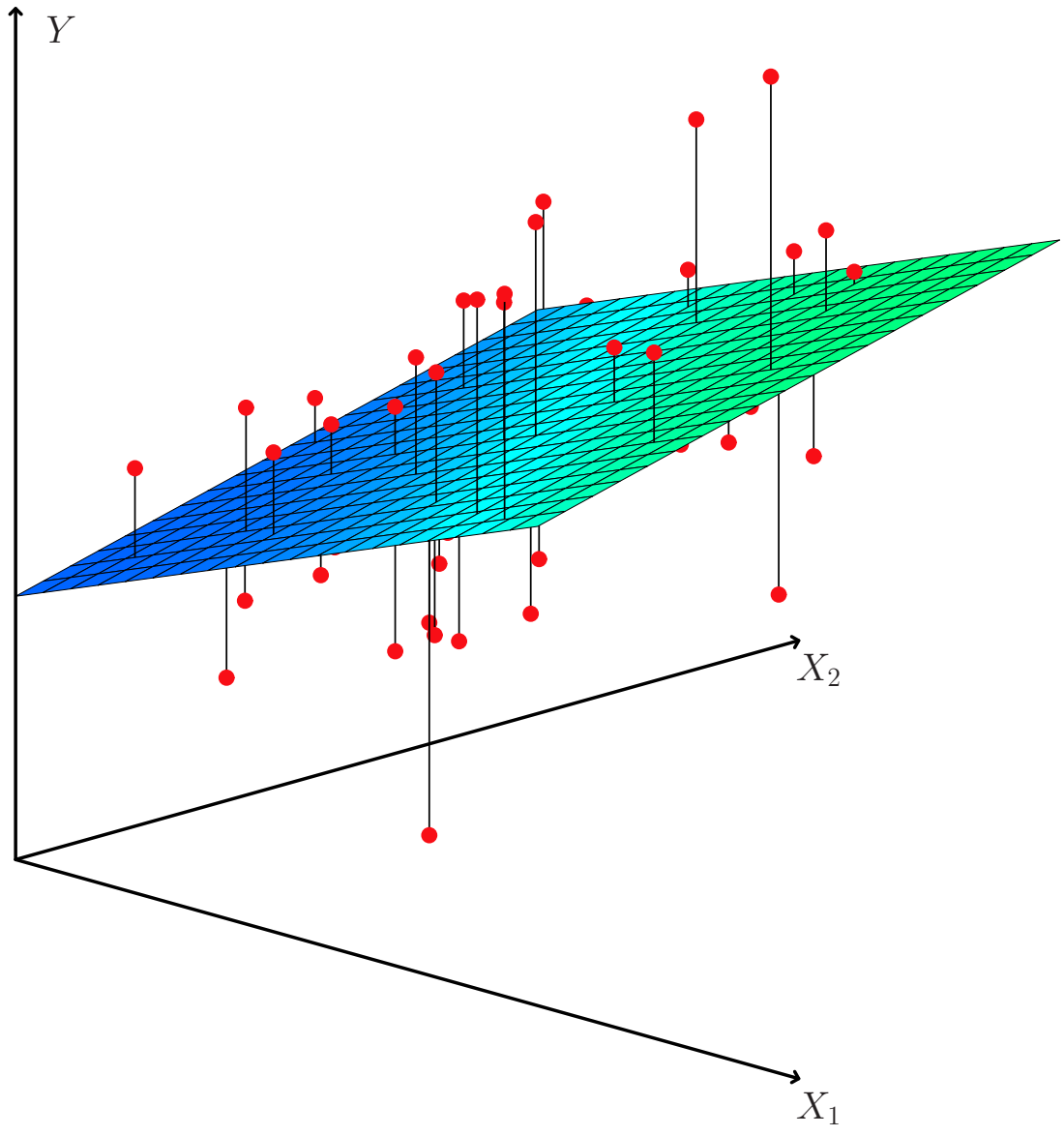
$$\texttt{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})^2.$$

- This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.
- Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$. Then the LS estimate $\hat{\boldsymbol{\beta}}$ can be neatly written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

where recall

$$\mathbf{X} = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{matrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{matrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{matrix}, \quad \mathbf{y} = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix}.$$


Y
X_2
X_1

Results for Advertising Data

Estimation results for multiple regression:

	Coefficient	Std.Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Estimation results for three simple linear regressions.

	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

	Coefficient	Std.Error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
Radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std.Error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

Explanation by Predictor Correlations

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- The correlation between radio and newspaper is 0.35: a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising. Now suppose that the multiple regression is correct and newspaper advertising has no direct impact on sales, but radio advertising does increase sales. Then in markets where we spend more on radio our sales will tend to be higher, we also tend to spend more on newspaper advertising in those same markets.
- Hence, in a simple linear regression which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising does not actually affect sales. So newspaper sales are a surrogate for radio advertising; newspaper gets "credit" for the effect of radio on sales.

The Woes of (interpreting) Regression Coefficients

"Data Analysis and Regression" Mosteller and Tukey 1977

- a regression coefficient β_j estimates the expected change in Y per unit change in X_j, with all other predictors held fixed. But predictors usually change together!
- $Y =$ number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $\hat{Y} = b_0 + 0.50W - 0.10H$. How do we interpret $\hat{\beta}_2 < 0$?
- Consider an absurd example to illustrate the point. Regressing shark attacks on ice cream sales for data collected at a given beach community over a period of time would show a positive relationship. Of course no one (yet) has suggested that ice creams should be banned at beaches to reduce shark attacks. In reality, higher temperatures cause more people to visit the beach, which in turn results in more ice cream sales and more shark attacks. A multiple regression of attacks versus ice cream sales and temperature reveals that, as intuition implies, the former predictor is no longer significant after adjusting for temperature.

Interpreting Regression Coefficients

- The ideal scenario is when the predictors are uncorrelated — a balanced design:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as "a unit change in X_j is associated with β_j change in Y, while all the other variables stay fixed", are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous — when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data.

Some Important Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ 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?

Is at Least One Predictor Useful?

For the first question, we test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_A : \text{ at least one } \beta_j \text{ is non-zero.}$$

We can use the F-statistic

$$F = \frac{(\texttt{TSS} - \texttt{RSS})/p}{\texttt{RSS}/(n-p-1)} \sim F_{p,n-p-1}.$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Compare F-Statistic to t-Statistics

- It seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response: this logic is flawed.
- Consider an example in which $p = 100$ and $\beta_1 = \cdots = \beta_p = 0$, so no variable is truly associated with the response. In this situation, about 5% of the p-values associated with each variable will be below 0.05 by chance. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response.
- Hence, if we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship.
- However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors.

Assessing the Overall Accuracy

- The quality of a LR fit is typically assessed using the residual standard error (RSE) and the adjusted R^2 statistic.
- Recall: RSS is $\texttt{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. RSE is defined as

$$\texttt{RSE} = \sqrt{\frac{1}{n-p-1}\texttt{RSS}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\texttt{TSS} - \texttt{RSS}}{\texttt{TSS}} = 1 - \frac{\texttt{RSS}}{\texttt{TSS}},$$

  where $\texttt{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares. But R^2 always increases when more predictors are used.
- It is common to use the adjusted R^2:

$$\bar{R}^2 = 1 - \frac{\texttt{RSS}/(n-p-1)}{\texttt{TSS}/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1}.$$

Deciding on the Important Variables

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models! Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

Forward Selection

- Begin with the null model — a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward Selection

- Start with all variables in the model.
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Uncertainty Associated with Prediction

Once the multiple regression model is fit, we can apply

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{1}$$

in order to predict the response Y on the basis of a set of values for the predictors $X_1, X_2, \ldots, X_p$. However, there are three sorts of uncertainty associated with this prediction.

(i) The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, \ldots, \beta_p$. That is, the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

is only an estimate for the *true population regression plane*

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The inaccuracy in the coefficient estimates is related to the *reducible error*. We can compute a *confidence interval* in order to determine how close $\hat{Y}$ will be to $f(X)$.

Uncertainty Associated with Prediction

(ii) In practice assuming a linear model for $f(X)$ is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call *model bias*. So when we use a linear model, we are in fact estimating the best linear approximation to the true surface. However, here we will ignore this discrepancy, and operate as if the linear model were correct.

(iii) Even if $f(X)$ follows the linear model and we know the true values for $\beta_0, \beta_1, \ldots, \beta_p$, the response value cannot be predicted perfectly because of the random error ϵ in the model. We use prediction intervals to check the difference between Y and $\hat{Y}$. Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

Uncertainty Associated with Prediction

We use a confidence interval to quantify the uncertainty surrounding the **average** sales over many cities. For example, given that $\$100,000$ is spent on TV advertising and $\$20,000$ is spent on radio advertising in each city, the 95% confidence interval is $[10,985, 11,528]$. We interpret this to mean that 95% of intervals of this form will contain the true value of $f(X)$.

On the other hand, a prediction interval can be used to quantify the uncertainty surrounding sales for a particular city. Given that $\$100,000$ is spent on TV advertising and $\$20,000$ is spent on radio advertising in that city the 95% prediction interval is $[7,930, 14,580]$. We interpret this to mean that 95% of intervals of this form will contain the true value of Y for this city. Note that both intervals are centered at $11,256$, but that the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about sales for a given city in comparison to the average sales over many locations.

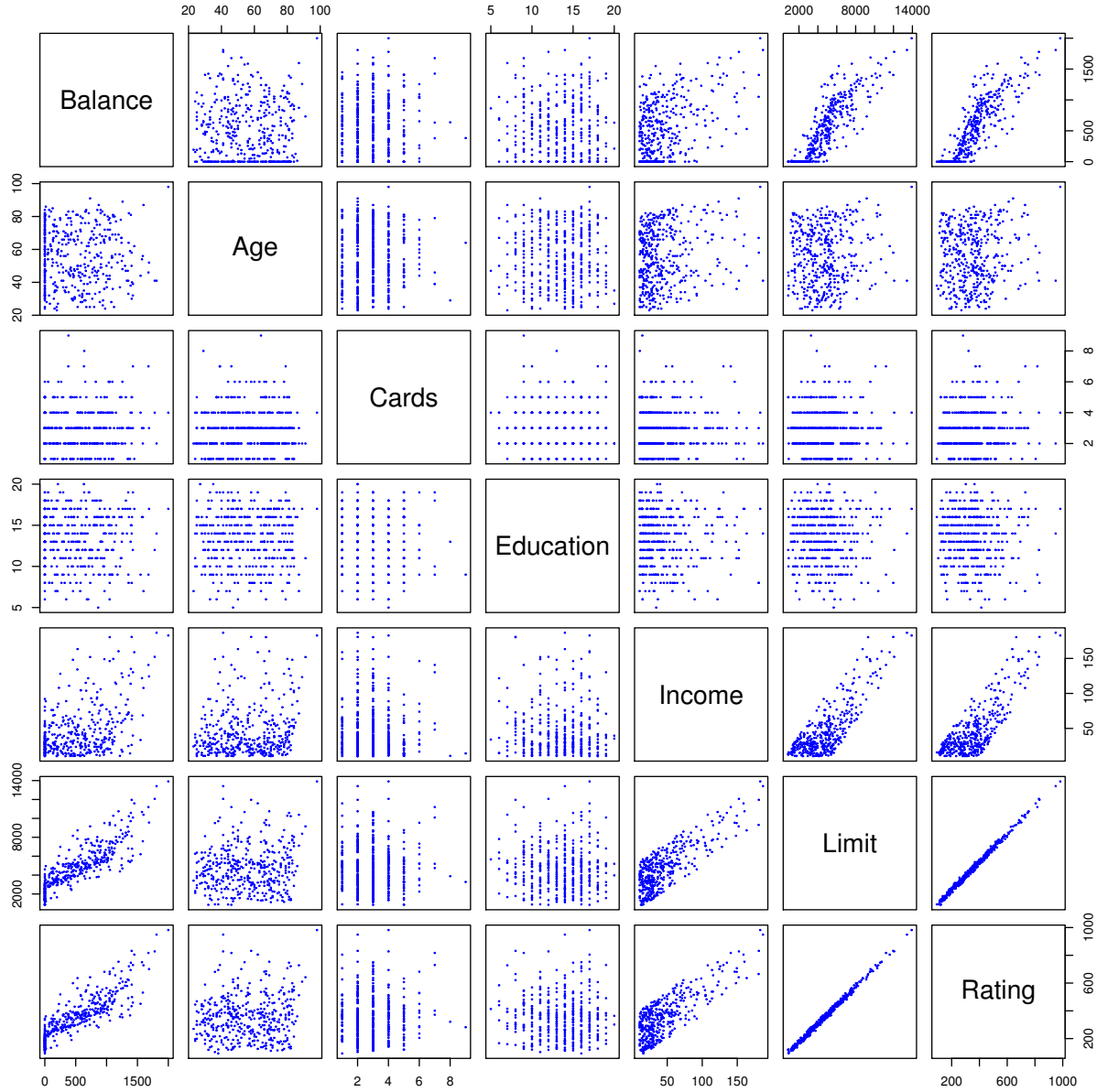
Overview

Qualitative Predictors

Some predictors are not quantitative but are qualitative, taking a discrete set of values. These are also called categorical predictors or factor variables.

- The Credit data set displayed in next slide records balance (average credit card debt for a number of individuals) as well as several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating).
- In addition to the 7 quantitative variables shown, there are four qualitative variables: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American (AA) or Asian).

Credit Card Data



Example: Males versus Females

We investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}.$$

- Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}.$$

- Interpretation: β_0 – the average credit card balance among males; $\beta_0 + \beta_1$ – average credit card balance among females, and β_1 – average difference in credit card balance between females and males.
- Fitted results for the gender model:

	Coefficient	Std.Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Qualitative Predictors: > 2 Levels I

- With more than two levels, we create additional dummy variables. For example, for the ethnicity variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}.$$

and the second is

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}.$$

- Then both of these variables can be used in the regression model.

Qualitative Predictors: > 2 Levels I

- The regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA} \end{cases}.$$

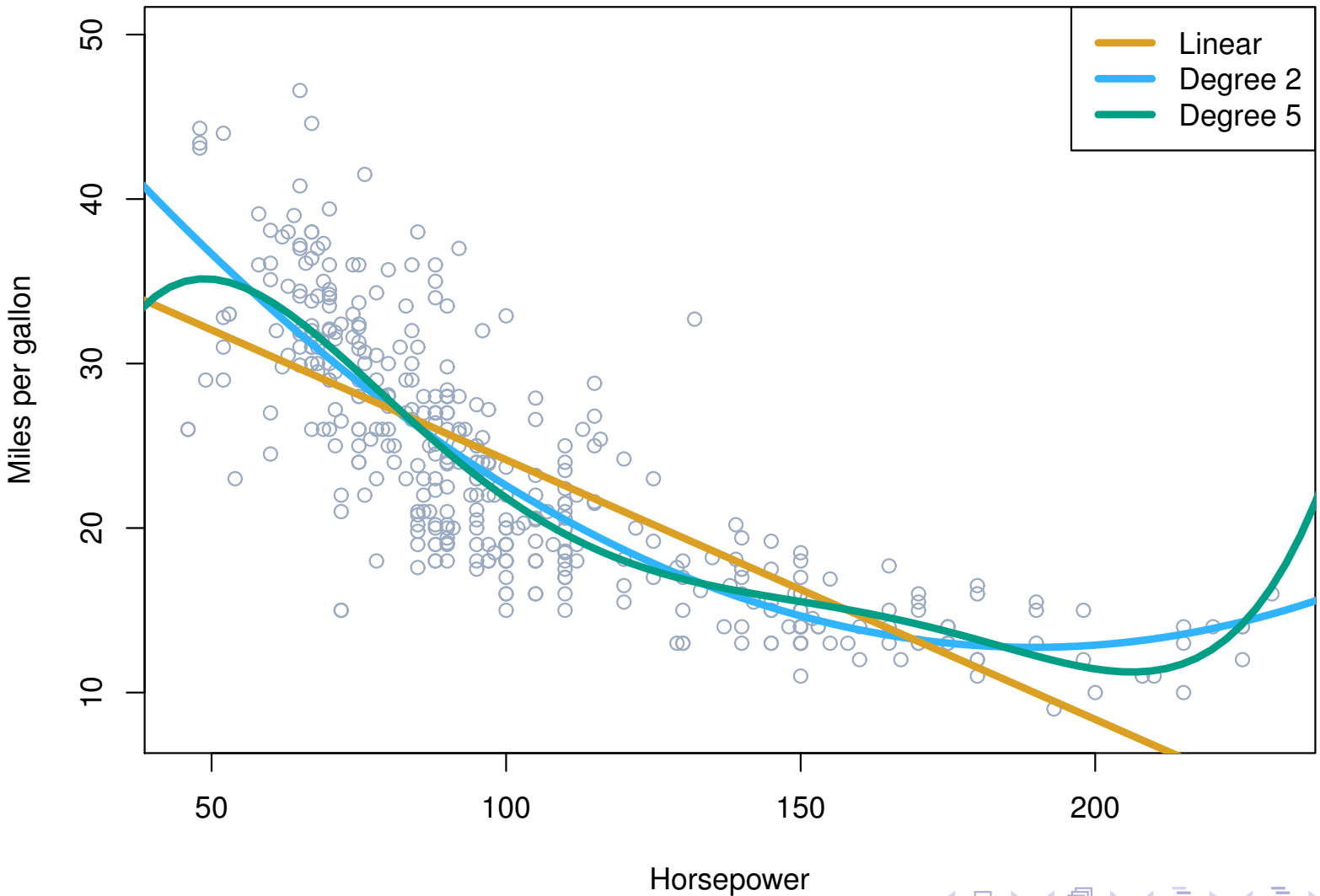
- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the baseline.
- Interpretation: β_0 – average cc balance for African Americans, β_1 – the difference in the average balance between the Asian and African American categories, and β_2 – the difference in the average balance between the Caucasian and African American categories.

Results for ethnicity:

	Coefficient	Std.Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	$<$0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Non-linear effects of predictors

The Auto dataset contains the mpg (gas mileage in miles per gallon) versus horsepower is shown for a number of cars.



Polynomial Model and Fitted Results

There is a pronounced relationship between mpg and horsepower, but it seems clear that this relationship is in fact non-linear: the data suggest a curved relationship. We can use a quadratic model to fit the data:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

We can still treat this model as a multiple linear regression model with $X_1 = \texttt{horsepower}$, and $X_2 = \texttt{horsepower}^2$. The previous figure suggests that this model fits the data better than the simple LR.

	Coefficient	Std.Error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	2 0.0012	0.0001	10.1	< 0.0001

Similarly, we can include higher order terms $X_3 = \texttt{horsepower}^3$, $X_4 = \texttt{horsepower}^4$, $X_5 = \texttt{horsepower}^5 \cdots\cdots$. This approach for extending the linear model to accommodate non-linear relationships is known as polynomial regression, since we have included polynomial functions of the predictors in the regression model.