CS 397: Topics in Computer Science–Probability & Statistics

Dr. Francis Parisi

Pace University

Early Summer 2019

Regression & Correlation

Linear regression is useful for predicting a quantitative response

Linear regression is useful for predicting a quantitative response

Although it has been around for a very long time, it is still one of the most widely used statistical learning methods

Probability & Statistics for Computer Science

Several questions we can answer with linear regression are:

Several questions we can answer with linear regression are:

- Is there a relationship between two or more variables?

Probability & Statistics for Computer Science

Several questions we can answer with linear regression are:

- Is there a relationship between two or more variables?
- How strong is the relationship?

Probability & Statistics for Computer Science

Several questions we can answer with linear regression are:

- Is there a relationship between two or more variables?
- How strong is the relationship?
- What are the effects of the different variables?

Probability & Statistics for Computer Science

Several questions we can answer with linear regression are:

- Is there a relationship between two or more variables?
- How strong is the relationship?
- What are the effects of the different variables?
- Can we make accurate predictions?
- Is the relationship linear?
- Is there synergy among the variables?

Simple Linear Regression In the simple linear regression (SLR) model there is one independent variable and has the form

$$Y \approx \beta_0 + \beta_1 X$$

Simple Linear Regression In the simple linear regression (SLR) model there is one independent variable and has the form

$$Y \approx \beta_0 + \beta_1 X$$

β_0 is the intercept

Simple Linear Regression In the simple linear regression (SLR) model there is one independent variable and has the form

$$Y \approx \beta_0 + \beta_1 X$$

β_0 is the intercept

β_1 is the slope of the regression line

Simple Linear Regression In the simple linear regression (SLR) model there is one independent variable and has the form

$$Y \approx \beta_0 + \beta_1 X$$

β_0 is the intercept

β_1 is the slope of the regression line

β_0, β_1 are coefficients, or model parameters

Model training results in the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Model training results in the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

We can use these estimates to predict the response value for a given input

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

where $\hat{y} = \mathbb{E}[Y|X = x]$

- The training data make up n ordered pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- The training data make up n ordered pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- We use the data to find the line that is closest to the data

- The training data make up n ordered pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- We use the data to find the line that is closest to the data
- SLR is based on the idea of least squares – find the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that describe the line that minimizes the squared errors

$\hat{y}_i$ is the predicted value for the i-th observation

$\hat{y}_i$ is the predicted value for the i-th observation

$e_i = y_i - \hat{y}_i$ is the i-th residual (error) term

$\hat{y}_i$ is the predicted value for the i-th observation

$e_i = y_i - \hat{y}_i$ is the i-th residual (error) term

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Probability & Statistics for Computer Science

$\hat{y}_i$ is the predicted value for the i-th observation

$e_i = y_i - \hat{y}_i$ is the i-th residual (error) term

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least squares minimizes RSS

Probability & Statistics for Computer Science

Using calculus we can find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize RSS are

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

and

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

How good are the coefficient estimates?

How good are the coefficient estimates?

- The true relationship between X and Y is $Y = f(X) + \epsilon$ for some unknown function f

How good are the coefficient estimates?

- The true relationship between X and Y is $Y = f(X) + \epsilon$ for some unknown function f
- $\hat{\beta}_0$, the intercept, is the value of Y when $X = 0$

How good are the coefficient estimates?

- The true relationship between X and Y is $Y = f(X) + \epsilon$ for some unknown function f
- $\hat{\beta}_0$, the intercept, is the value of Y when $X = 0$ this is not always interpretable

How good are the coefficient estimates?

- The true relationship between X and Y is $Y = f(X) + \epsilon$ for some unknown function f
- $\hat{\beta}_0$, the intercept, is the value of Y when $X = 0$ this is not always interpretable
- $\hat{\beta}_1$, the slope, is the average change in Y for a one-unit change in X

How good are the coefficient estimates?

- The true relationship between X and Y is $Y = f(X) + \epsilon$ for some unknown function f
- $\hat{\beta}_0$, the intercept, is the value of Y when $X = 0$ this is not always interpretable
- $\hat{\beta}_1$, the slope, is the average change in Y for a one-unit change in X
- ϵ is the error term

The population regression line is

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

and the least squares regression line is

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

The population regression line is

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

and the least squares regression line is

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased* estimates

Probability & Statistics for Computer Science

The population regression line is

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

and the least squares regression line is

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased* estimates – if we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ repeatedly over many data sets drawn from the same population then $\hat{\beta}_0 \to \beta_0$ and $\hat{\beta}_1 \to \beta_1$

If we estimate the population mean as $\hat{\mu}$ then

$$\mathrm{Var}\left(\hat{\mu}\right) = \frac{\sigma^2}{n} = SE(\hat{\mu}) \tag{14.1}$$

where $\sigma^2 = \mathrm{Var}\left(\epsilon\right)$

If we estimate the population mean as $\hat{\mu}$ then

$$\operatorname{Var}(\hat{\mu}) = \frac{\sigma^2}{n} = SE(\hat{\mu}) \tag{14.1}$$

where $\sigma^2 = \operatorname{Var}(\epsilon)$

The standard error $SE(\hat{\mu})$ gives the average amount by which $\hat{\mu}$ differs from μ

Additionally,

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

and

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

In general we don't know σ but we estimate it as the residual standard error

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

In general we don't know σ but we estimate it as the residual standard error

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

We can use RSE to find confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$

In general we don't know σ but we estimate it as the residual standard error

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

We can use RSE to find confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$

95% confidence intervals are $\hat{\beta}_0 \pm 1.96 \times SE(\hat{\beta}_0)$ and $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$

In general we don't know σ but we estimate it as the residual standard error

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

We can use RSE to find confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$

95% confidence intervals are $\hat{\beta}_0 \pm 1.96 \times SE(\hat{\beta}_0)$ and $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$

What does this mean?

In general we don't know σ but we estimate it as the residual standard error

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

We can use RSE to find confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$

95% confidence intervals are $\hat{\beta}_0 \pm 1.96 \times SE(\hat{\beta}_0)$ and $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$

What does this mean? There's a 95% chance that the interval contains the true β_i

- We use the RSE to conduct hypothesis testing

- We use the RSE to conduct hypothesis testing
- In linear regression we have

H_0 : There is no relationship between X and Y

- We use the RSE to conduct hypothesis testing
- In linear regression we have

$$\begin{aligned} H_0 &: \quad \text{There is no relationship between } X \text{ and } Y \\ H_a &: \quad \text{There is some relationship between } X \text{ and } Y \end{aligned}$$

- We use the RSE to conduct hypothesis testing
- In linear regression we have

$$
\begin{aligned}
H_0 &: \quad \text{There is no relationship between } X \text{ and } Y \\
H_a &: \quad \text{There is some relationship between } X \text{ and } Y \\
H_0 &: \quad \beta_1 = 0, \quad H_a : \beta_1 \neq 0
\end{aligned}
$$

- We use the RSE to conduct hypothesis testing
- In linear regression we have

$$
\begin{aligned}
H_0 &: \quad \text{There is no relationship between } X \text{ and } Y \\
H_a &: \quad \text{There is some relationship between } X \text{ and } Y \\
H_0 &: \quad \beta_1 = 0, \quad H_a : \beta_1 \neq 0
\end{aligned}
$$

- Does the estimate $\hat{\beta}_1$ differ significantly from zero?

- This depends on the value of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$

- This depends on the value of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$
- Proceed by computing the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

when $H_0 : \beta_1 = 0$

- This depends on the value of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$
- Proceed by computing the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

  when $H_0 : \beta_1 = 0$
- This statistic has a t-distribution with $n-2$ degrees of freedom

- This depends on the value of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$
- Proceed by computing the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

when $H_0 : \beta_1 = 0$

- This statistic has a t-distribution with $n - 2$ degrees of freedom
- The p-value associated with this t-statistic helps us make a decision about H_0

Definition

p-value is the probability of observing a value of $|t|$ or greater (absolute value) when H_0 is true

- A small p-value suggests there is enough evidence to reject H_0 and conclude $\beta_1 \neq 0$

Definition

p-value is the probability of observing a value of $|t|$ or greater (absolute value) when H_0 is true

- A small p-value suggests there is enough evidence to reject H_0 and conclude $\beta_1 \neq 0$
- In other words, with a small p-value it is unlikely that H_0 is true

`Advertising` data contains the amount spent on different forms of advertising and the amount of sales

`Advertising` data contains the amount spent on different forms of advertising and the amount of sales

Table 14.1: Least squares coefficients for Regressing Sales on TV ads from 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

Advertising data contains the amount spent on different forms of advertising and the amount of sales

Table 14.1: Least squares coefficients for Regressing Sales on TV ads from 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

The results show that there is a statistically significant relationship between money spent of **TV** ads and **Sales**

- If we reject H_0 we must ask 'how good is the model?'

- If we reject H_0 we must ask 'how good is the model?'
- We typically use RSE and R^2 to assess the quality of the fit

- If we reject H_0 we must ask 'how good is the model?'
- We typically use RSE and R^2 to assess the quality of the fit

Recall $RSE = \sqrt{\frac{1}{n-2}RSS}$, and we define the **Total Sum of Squares**

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

and

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

Probability & Statistics for Computer Science

- RSE is an estimate of the standard deviation of ϵ

- RSE is an estimate of the standard deviation of ϵ – the average amount by which the response will deviate from the true regression

- RSE is an estimate of the standard deviation of ϵ – the average amount by which the response will deviate from the true regression
- While RSE is a measure of goodness of fit it is hard to tell at what value is the fit good or bad

- RSE is an estimate of the standard deviation of ϵ – the average amount by which the response will deviate from the true regression
- While RSE is a measure of goodness of fit it is hard to tell at what value is the fit good or bad
- R^2 is the proportion of variability in Y explained by X in the model

- RSE is an estimate of the standard deviation of ϵ – the average amount by which the response will deviate from the true regression
- While RSE is a measure of goodness of fit it is hard to tell at what value is the fit good or bad
- R^2 is the proportion of variability in Y explained by X in the model
- $R^2 \in [0, 1]$ so R^2 from different models are comparable

- R^2 is a measure of the linear relationship between X and Y

- R^2 is a measure of the linear relationship between X and Y
- Correlation is also a measure of the linear relationship between X and Y

- R^2 is a measure of the linear relationship between X and Y
- Correlation is also a measure of the linear relationship between X and Y

$$r = \mathrm{Cor}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

- R^2 is a measure of the linear relationship between X and Y
- Correlation is also a measure of the linear relationship between X and Y

$$r = \mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

- For SLR $R^2 = r^2$, however this does not hold for Multiple Linear Regression

- R^2 is a measure of the linear relationship between X and Y
- Correlation is also a measure of the linear relationship between X and Y

$$r = \mathrm{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- For SLR $R^2 = r^2$, however this does not hold for Multiple Linear Regression – thus R^2 is the measure we'll use

- Correlation analysis attempts to measure the strength of the relationship between two variables by means of a single number called a correlation coefficient

- Correlation analysis attempts to measure the strength of the relationship between two variables by means of a single number called a correlation coefficient
- Above we looked at the sample correlation r which is an estimate of the population correlation, ρ

Exercise

Compute and interpret the correlation coefficient for the following grades of 6 students selected at random:

Math Grade	70	92	80	74	65	83
English Grade	74	84	63	87	78	90

Multiple Linear Regression

Extends the SLR model to one with two or more independent variables

Extends the SLR model to one with two or more independent variables

`Advertising` example: **`sales`** $\sim$ **`TV`**, what about the other variables, **`radio`** and **`newspaper`**?

Extends the SLR model to one with two or more independent variables

`Advertising` example: `sales` $\sim$ `TV`, what about the other variables, `radio` and `newspaper`?

One way to deal with this is to make three regression models, but that isn't a good solution

Extends the SLR model to one with two or more independent variables

`Advertising` example: **`sales`** $\sim$ **`TV`**, what about the other variables, **`radio`** and **`newspaper`**?

One way to deal with this is to make three regression models, but that isn't a good solution

- With 3 models we can't make a single prediction for **`sales`**

Extends the SLR model to one with two or more independent variables

`Advertising` example: **`sales`** $\sim$ **`TV`**, what about the other variables, **`radio`** and **`newspaper`**?

One way to deal with this is to make three regression models, but that isn't a good solution

- With 3 models we can't make a single prediction for **`sales`**
- Coefficient estimates ignore the effects of the other variables

Extends the SLR model to one with two or more independent variables

`Advertising` example: **`sales`** $\sim$ **`TV`**, what about the other variables, **`radio`** and **`newspaper`**?

One way to deal with this is to make three regression models, but that isn't a good solution

- With 3 models we can't make a single prediction for **`sales`**
- Coefficient estimates ignore the effects of the other variables

Instead we fit a MLR model using all the variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{14.19}$$

Estimating the Coefficients

Given $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ we can make predictions via

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

Given $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ we can make predictions via

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

Like SLR we find the $\hat{\beta}$'s that minimize the RSS

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y})^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p)^2 \end{aligned}$$

Given $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ we can make predictions via

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

Like SLR we find the $\hat{\beta}$'s that minimize the RSS

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y})^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p)^2 \end{aligned}$$

In matrix notation the coefficients are

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

Let's compare ...

Table 14.3: Simple regression of **sales** on **radio** (top) and **sales** on **newspaper** (bottom)

	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	< 0.0001

Let's compare ...

Table 14.3: Simple regression of **sales** on **radio** (top) and **sales** on **newspaper** (bottom)

	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	< 0.0001

Both **radio** and **newspaper** are statistically significant

Table 14.4: Multiple regression of **sales** on **TV, radio, newspaper**

	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

Table 14.4: Multiple regression of **`sales`** on **`TV, radio, newspaper`**

	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

When all three predictors are in the model, **`newspaper`** is no longer significant *given* **`TV`** and **`radio`** are already included

1. *Is there a relationship between the response and the independent variables?*

1. *Is there a relationship between the response and the independent variables?*

In the SLR case we test the hypothesis that $\beta_1 = 0$, i.e., no relationship

1. *Is there a relationship between the response and the independent variables?*

In the SLR case we test the hypothesis that $\beta_1 = 0$, i.e., no relationship

In MLR with p predictors we test

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

1. *Is there a relationship between the response and the independent variables?*

In the SLR case we test the hypothesis that $\beta_1 = 0$, i.e., no relationship

In MLR with p predictors we test

$$
\begin{aligned}
H_0: &\quad \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\
H_a: &\quad \text{at least one is not zero}
\end{aligned}
$$

1. *Is there a relationship between the response and the independent variables?*

In the SLR case we test the hypothesis that $\beta_1 = 0$, i.e., no relationship

In MLR with p predictors we test

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_a &: \quad \text{at least one is not zero} \end{aligned}$$

The F-statistic in the MLR case is

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{14.23}$$

Linear Regression

If the linear model assumptions are correct and H_0 is true then

$$\mathbb{E}[RSS/(n-p-1)] = \sigma^2$$

and

$$\mathbb{E}[(TSS - RSS)/p] = \sigma^2$$

so we expect $F \approx 1$

Linear Regression

If the linear model assumptions are correct and H_0 is true then

$$\mathbb{E}[RSS/(n-p-1)] = \sigma^2$$

and

$$\mathbb{E}[(TSS - RSS)/p] = \sigma^2$$

so we expect $F \approx 1$

If H_a is true then

$$\mathbb{E}[(TSS - RSS)/p] > \sigma^2$$

so we expect $F > 1$

A large F-statistic suggests at least one $\beta_i \neq 0$

A large F-statistic suggests at least one $\beta_i \neq 0$

F depends on n and p so sometimes it is unclear how large F needs to be to reject H_0

A large F-statistic suggests at least one $\beta_i \neq 0$

F depends on n and p so sometimes it is unclear how large F needs to be to reject H_0

We find the p-value associated with the F-statistic and make a decision

In (14.23) we're testing all β's $= 0$

In (14.23) we're testing all β's $= 0$

Sometimes we want to test some subset q of the β's $= 0$

In (14.23) we're testing all β's $= 0$

Sometimes we want to test some subset q of the β's $= 0$

Under this scenario H_0 becomes

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

In (14.23) we're testing all β's $= 0$

Sometimes we want to test some subset q of the β's $= 0$

Under this scenario H_0 becomes

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

Here for convenience we put the q variables of interest at the end of the list

In (14.23) we're testing all β's $= 0$

Sometimes we want to test some subset q of the β's $= 0$

Under this scenario H_0 becomes

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

Here for convenience we put the q variables of interest at the end of the list

We then fit a models using all the variables except the last q and find RSS_0, the RSS for the reduced model

The F-statistic is now

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

The F-statistic is now

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

The R output we looked at earlier produces partial t-statistics for each variable *given* the other variables above it are in the model

The F-statistic is now

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

The R output we looked at earlier produces partial t-statistics for each variable *given* the other variables above it are in the model

Looking at the individual p-values to determine whether or not to reject H_0 is not enough

The F-statistic is now

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

The R output we looked at earlier produces partial t-statistics for each variable *given* the other variables above it are in the model

Looking at the individual p-values to determine whether or not to reject H_0 is not enough

Given a large number of variables, say $p = 100$ about 5% of the p-values will be below 0.05 just by chance

The F-statistic is now

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

The R output we looked at earlier produces partial t-statistics for each variable *given* the other variables above it are in the model

Looking at the individual p-values to determine whether or not to reject H_0 is not enough

Given a large number of variables, say $p = 100$ about 5% of the p-values will be below 0.05 just by chance

We don't have this problem with F because the statistic adjusts for p

For large p *forward selection* helps (next section)

For large p *forward selection* helps (next section)

If $p > n$ we cannot fit a model

2. *Find the important variables*

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only, and (4) Both

How do we find the 'best' model?

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only, and (4) Both

How do we find the 'best' model?

- Mallows C_p

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only, and (4) Both

How do we find the 'best' model?

- Mallows C_p
- Akaike Information Criterion (AIC)

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only, and (4) Both

How do we find the 'best' model?

- Mallows C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only, and (4) Both

How do we find the 'best' model?

- Mallows C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Adjusted R^2

2. *Find the important variables*

Variable selection is determining which predictors are related to the response

Ideally consider all possible subsets – let's say $p = 2$
This gives 4 cases (anyone recall power sets?)

(1) No variables (2) X_1 only (3) X_2 only, and (4) Both

How do we find the 'best' model?

- Mallows C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Adjusted R^2

We will discuss these more in Chapter 6

Given some value for p there are 2^p possible models and as p grows it becomes infeasible to try them all

Probability & Statistics for Computer Science

Given some value for p there are 2^p possible models and as p grows it becomes infeasible to try them all

Three approaches to model building

Given some value for p there are 2^p possible models and as p grows it becomes infeasible to try them all

Three approaches to model building

Forward Selection: Start with a null model using β_0. Fit p individual (SLR) models and add the variable with the lowest RSS to the null model. Keep going until you cannot add more.

Probability & Statistics for Computer Science

Given some value for p there are 2^p possible models and as p grows it becomes infeasible to try them all

Three approaches to model building

Forward Selection: Start with a null model using β_0. Fit p individual (SLR) models and add the variable with the lowest RSS to the null model. Keep going until you cannot add more.

Backward Selection: Start with all the variables and remove the one with the highest p-value. Continue until there are no variables with a p-value $> \alpha$.

Given some value for p there are 2^p possible models and as p grows it becomes infeasible to try them all

Three approaches to model building

Forward Selection: Start with a null model using β_0. Fit p individual (SLR) models and add the variable with the lowest RSS to the null model. Keep going until you cannot add more.

Backward Selection: Start with all the variables and remove the one with the highest p-value. Continue until there are no variables with a p-value $> \alpha$.

Mixed Selection: Start with the null model and add variables like forward selection. If the p-value of a variable already in the model goes above α take it out. Continue until all the variables in the model have p-values less than α and adding more would have p-value $> \alpha$

- Backward selection cannot be used if $p > n$

- Backward selection cannot be used if $p > n$
- Forward selection can always be used, but may end up leaving out variables

- Backward selection cannot be used if $p > n$
- Forward selection can always be used, but may end up leaving out variables
- Mixed selection always works and doesn't have these issues

Probability & Statistics for Computer Science

3. *Assessing Model Fit*

3. *Assessing Model Fit*

- Recall that RSE and R^2 are the two most common measures

3. *Assessing Model Fit*

- Recall that RSE and R^2 are the two most common measures
- For SLR $R^2 = \mathrm{Cor}(X, Y)^2$ and in MLR $R^2 = \mathrm{Cor}(Y, \hat{Y})$

3. *Assessing Model Fit*

- Recall that RSE and R^2 are the two most common measures
- For SLR $R^2 = \mathrm{Cor}(X, Y)^2$ and in MLR $R^2 = \mathrm{Cor}(Y, \hat{Y})$
- The fitted linear model maximizes this correlation among all possible models

3. *Assessing Model Fit*

- Recall that RSE and R^2 are the two most common measures
- For SLR $R^2 = \mathrm{Cor}(X, Y)^2$ and in MLR $R^2 = \mathrm{Cor}(Y, \hat{Y})$
- The fitted linear model maximizes this correlation among all possible models

Be aware that R^2 alays increases as more variables are added even if they are only weakly related to the response

3. *Assessing Model Fit*

- Recall that RSE and R^2 are the two most common measures
- For SLR $R^2 = \mathrm{Cor}(X, Y)^2$ and in MLR $R^2 = \mathrm{Cor}(Y, \hat{Y})$
- The fitted linear model maximizes this correlation among all possible models

Be aware that R^2 alays increases as more variables are added even if they are only weakly related to the response – adjusted R^2 addresses this (later)

3. *Assessing Model Fit*

- Recall that RSE and R^2 are the two most common measures
- For SLR $R^2 = \text{Cor}(X, Y)^2$ and in MLR $R^2 = \text{Cor}(Y, \hat{Y})$
- The fitted linear model maximizes this correlation among all possible models

Be aware that R^2 alays increases as more variables are added even if they are only weakly related to the response – adjusted R^2 addresses this (later)

RSE in MLR is

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

which reduces to the equation we for SLR when $p = 1$

Probability & Statistics for Computer Science

Visualization

Visualization

Graphs are essential and can highlight problems not detectable with numerical statistics

Visualization

Graphs are essential and can highlight problems not detectable with numerical statistics

The residuals in the following graph suggest a non-linear relationship which is probably due to an interaction between variables (more later)

Probability & Statistics for Computer Science

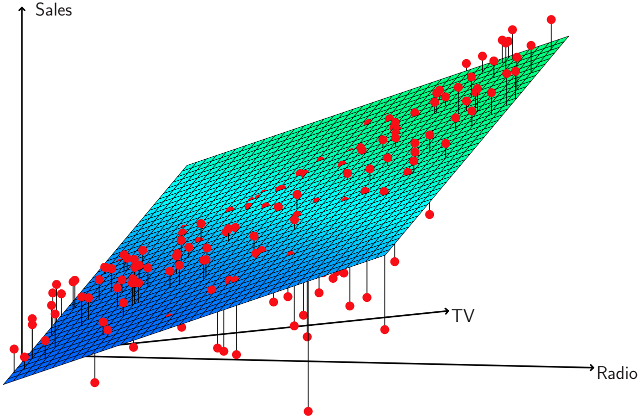


Figure 14.5: Response surface **Advertising** data **sales**∼**TV**+**radio**

4. *Making Predictions*

4. *Making Predictions*

We can use our model to predict the reponse Y given a set of predictors $X_1, X_2, \ldots, X_p$

4. *Making Predictions*

We can use our model to predict the reponse Y given a set of predictors $X_1, X_2, \ldots, X_p$

Uncertainty associated with prediction

4. *Making Predictions*

We can use our model to predict the reponse Y given a set of predictors $X_1, X_2, \ldots, X_p$

Uncertainty associated with prediction

(a) $\hat{\beta}$'s are estimates for the true β's

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p \text{ vs. the true}$$
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
$$f(X) - \hat{Y} : \text{Reducible error}$$

4. *Making Predictions*

We can use our model to predict the reponse Y given a set of predictors $X_1, X_2, \ldots, X_p$

Uncertainty associated with prediction

(a) $\hat{\beta}$'s are estimates for the true β's

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p \text{ vs. the true}$$

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

$$f(X) - \hat{Y} : \text{Reducible error}$$

A *confidence interval* quantifies the uncertainty around how close $\hat{Y}$ is to $f(X)$

4. *Making Predictions*

We can use our model to predict the reponse Y given a set of predictors $X_1, X_2, \dots, X_p$

Uncertainty associated with prediction

(a) $\hat{\beta}$'s are estimates for the true β's

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p \text{ vs. the true}$$
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
$$f(X) - \hat{Y} : \text{Reducible error}$$

A *confidence interval* quantifies the uncertainty around how close $\hat{Y}$ is to $f(X)$

(b) Model bias – assuming reality is correctly modeled by our linear assumption

(b) Model bias – assuming reality is correctly modeled by our linear assumption

(c) Irreducible error which cannot be eliminated

(b) Model bias – assuming reality is correctly modeled by our linear assumption
(c) Irreducible error which cannot be eliminated

Prediction intervals capture reducible error and irreducible error and are always wider than confidence intervals

(b) Model bias – assuming reality is correctly modeled by our linear assumption
(c) Irreducible error which cannot be eliminated

Prediction intervals capture reducible error and irreducible error and are always wider than confidence intervals

Confidence intervals quantify the uncertainty around the average predicted value given X

(b) Model bias – assuming reality is correctly modeled by our linear assumption

(c) Irreducible error which cannot be eliminated

Prediction intervals capture reducible error and irreducible error and are always wider than confidence intervals

Confidence intervals quantify the uncertainty around the average predicted value given X

Prediction intervals quantify the uncertainty around a specific predicted value given X

- Predictor variables are not always quantitative and some may be qualitative

- Predictor variables are not always quantitative and some may be qualitative
- The `Credit` data set has a number of quantitative variables and four qualitative ones: `gender`, `student`, (marital) `status`, and `ethnicity`

- Predictor variables are not always quantitative and some may be qualitative
- The **`Credit`** data set has a number of quantitative variables and four qualitative ones: **`gender`**, **`student`**, (marital) **`status`**, and **`ethnicity`**

Qualitative variables are also known as categorical variables, or factors `gender` has two levels so we use a "dummy" variable

$$x_i = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

`gender` has two levels so we use a "dummy" variable

$$x_i = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

In a regression model this becomes

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if female} \\ \beta_0 + \epsilon & \text{if male} \end{cases}$$

gender has two levels so we use a “dummy” variable

$$x_i = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

In a regression model this becomes

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if female} \\ \beta_0 + \epsilon & \text{if male} \end{cases}$$

For the **Credit** data, β_0 is the average card balance for males, and $\beta_0 + \beta_1$ is the average card balance for females

Here we list the regression model output and notice that there is no statistically significant evidence that card balance varies by gender (p-value $= 0.6690$)

Table 14.7: Least squares estimates for `balance` on `gender`

	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

A 0/1 coding is not the only way to code a dummy variable

A 0/1 coding is not the only way to code a dummy variable

$$x_i = \begin{cases} 1 & \text{if female} \\ -1 & \text{if male} \end{cases}$$

works just as well

A 0/1 coding is not the only way to code a dummy variable

$$x_i = \begin{cases} 1 & \text{if female} \\ -1 & \text{if male} \end{cases}$$

works just as well

The model becomes

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if female} \\ \beta_0 - \beta_1 + \epsilon & \text{if male} \end{cases}$$

As the number of levels grows so does the number of dummy variables

More Than 2 Levels

As the number of levels grows so does the number of dummy variables

For n levels we need $n - 1$ dummy variables

More Than 2 Levels

As the number of levels grows so does the number of dummy variables

For n levels we need $n - 1$ dummy variables

The **`ethnicity`** variable in the **`Credit`** data set has three levels African-American, Asian and Caucasian

As the number of levels grows so does the number of dummy variables

For n levels we need $n-1$ dummy variables

The `ethnicity` variable in the `Credit` data set has three levels African-American, Asian and Caucasian

$$x_{i1} = \begin{cases} 1 & \text{if Asian} \\ 0 & \text{if not Asian} \end{cases}$$

and

$$x_{i1} = \begin{cases} 1 & \text{if Caucasian} \\ 0 & \text{if not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Caucasian} \\ \beta_0 + \epsilon_i & \text{if African-American} \end{cases}$$

African-American in this example is the *baseline*

Table 3.8 shows the R output for this model, but the **ethnicity** variable is not statistically significant

Table 14.8: Least squares estimates for **balance** on **ethnicity**

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.646	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

Table 3.8 shows the R output for this model, but the **`ethnicity`** variable is not statistically significant

Table 14.8: Least squares estimates for **`balance`** on **`ethnicity`**

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.646	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Of course we can use quantitative and qualitative variables in the same MLR model

While the linear regression model works well in many real-world problems it makes restrictive assumptions

While the linear regression model works well in many real-world problems it makes restrictive assumptions – primarily that there are additive and linear relationships between the response and the predictor variables

While the linear regression model works well in many real-world problems it makes restrictive assumptions – primarily that there are additive and linear relationships between the response and the predictor variables

Sometimes the effects of the predictors are not independent as assumed

While the linear regression model works well in many real-world problems it makes restrictive assumptions – primarily that there are additive and linear relationships between the response and the predictor variables

Sometimes the effects of the predictors are not independent as assumed

When the change in one variable affects another we say there is an *interaction* effect (marketing people call this *synergy*)

Suppose we have a linear model with two variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Suppose we have a linear model with two variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

A unit change is say X_1 change Y by β_1 regardless of X_2

Suppose we have a linear model with two variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

A unit change is say X_1 change Y by β_1 regardless of X_2

We can extend this model to include an *interaction term* $X_1 X_2$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \qquad (14.31)$$

We can re-write (14.24) as

$$
\begin{aligned}
Y &= \beta_0 + (\beta_1 + \beta_3 X_2)X_1 + \beta_2 X_2 + \epsilon \\
&= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned} \tag{14.32}
$$

where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$

We can re-write (14.24) as

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned} \tag{14.32}$$

where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$

From (14.32) it's apparent that as X_2 changes so does $\tilde{\beta}_1$ which changes the effect of X_1 on Y, thus the interaction

Example

Suppose that we are interested in studying the productivity of a factory. We wish to predict the number of **`units`** *produced on the basis of the number of production* **`lines`** *and the total number of* **`workers`***. It seems likely that the effect of increasing the number of production* **`lines`** *will depend on the number of* **`workers`***, since if no* **`workers`** *are available to operate the* **`lines`***, then increasing the number of* **`lines`** *will not increase* **`production`***. This suggests that it would be appropriate to include an interaction term between* **`lines`** *and* **`workers`** *in a linear model to predict* **`units`***.*

Example

Say we get the following results after fitting the model

$$\begin{aligned}\texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} \\ &\qquad + 1.4 \times \texttt{lines} \times \texttt{workers} \\ &= 1.2 + (3.4 + 1.4 \times \texttt{workers}) \times \texttt{lines} + 0.22 \times \texttt{workers}\end{aligned}$$

Example

Say we get the following results after fitting the model

$$\begin{aligned}\texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} \\ &\quad + 1.4 \times \texttt{lines} \times \texttt{workers} \\ &= 1.2 + (3.4 + 1.4 \times \texttt{workers}) \times \texttt{lines} + 0.22 \times \texttt{workers}\end{aligned}$$

For additional `line` added the number of `units` goes up by $3.4 + 1.4 \times \texttt{workers}$

Example

Using the `Advertising` data from earlier, we regress `sales` on `radio`, `TV`, and the interaction between the two

Example

Using the **Advertising** data from earlier, we regress **sales** on **radio**, **TV**, and the interaction between the two

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon\end{aligned} \tag{14.33}$$

Example

Using the **Advertising** data from earlier, we regress **sales** on **radio**, **TV**, and the interaction between the two

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon\end{aligned} \tag{14.33}$$

Hereto we see that the effect of one variable (**TV**) is a function of another variable (**radio**)

The fitted model from (14.33) gives the following results

Table 14.9: Least squares estimates for **`sales`** on **`radio`**, **`TV`**, and their interaction.

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	6.7502	0.248	27.23	< 0.0001
`TV`	0.0191	0.002	12.70	< 0.0001
`radio`	0.0289	0.009	3.24	0.0014
`TV` $\times$ **`radio`**	0.0011	0.000	20.73	< 0.0001

The fitted model from (14.33) gives the following results

Table 14.9: Least squares estimates for **sales** on **radio**, **TV**, and their interaction.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV $\times$ **radio**	0.0011	0.000	20.73	< 0.0001

From the table we note that the main effects and interaction term are all statistically significant

From Table 14.9 we see all the variables, main effects and interaction, are significant

From Table 14.9 we see all the variables, main effects and interaction, are significant

That's not always the case

From Table 14.9 we see all the variables, main effects and interaction, are significant

That's not always the case – we may find the interaction significant but one or both main efffects not

From Table 14.9 we see all the variables, main effects and interaction, are significant

That's not always the case – we may find the interaction significant but one or both main efffects not

The *hierarchical principle* tells us if the interaction term is significant we should include the main effects even if one or both are not

Interactions also work with qualitative variables

Interactions also work with qualitative variables
From the **Credit** data we regress **balance** on **income** and the qualitative variable **student**

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if student} \\ 0 & \text{if not} \end{cases}$$

$$= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if student} \\ 0 & \text{if not} \end{cases}$$

It's possible that income has a different effect on students and non-students so we test this by including an interaction term

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \beta_2 \times \texttt{student}_i \\ &\quad + \beta_3 \times (\texttt{income}_i \times \texttt{student}_i) \\ &= \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_3 \times \texttt{income}_i & \text{if } \texttt{student} \\ 0 & \text{if not} \end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if } \texttt{s} \\ 0 & \text{if n} \end{cases}\end{aligned}$$

- With the additive assumption relaxed we can consider non-linear relationships

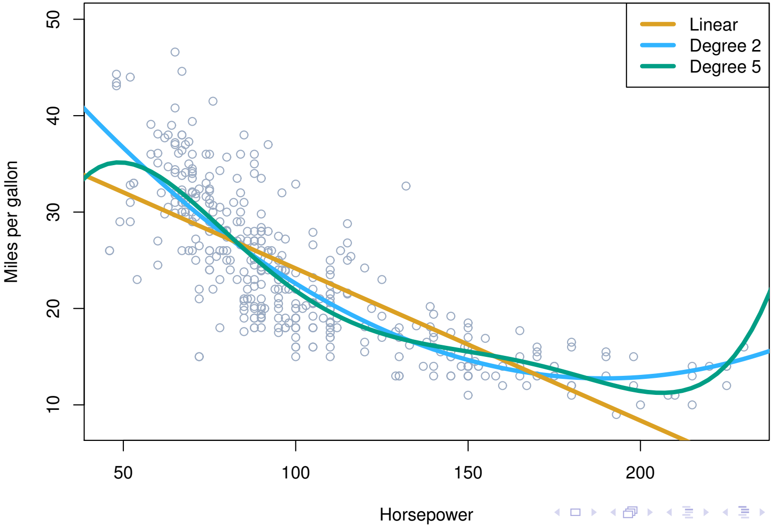
- With the additive assumption relaxed we can consider non-linear relationships
- The simplest non-linear extension of the linear model is a polynomial regression

- With the additive assumption relaxed we can consider non-linear relationships
- The simplest non-linear extension of the linear model is a polynomial regression
- Polynomial Regression a linear model with polynomial functions as predictors

- With the additive assumption relaxed we can consider non-linear relationships
- The simplest non-linear extension of the linear model is a polynomial regression
- Polynomial Regression a linear model with polynomial functions as predictors
- Using the **Auto** data we regress **mpg** on **horsepower** and plot the results for a linear model a second-degree polynomial and a fifth-degree polynomial on the next slide

Probability & Statistics for Computer Science

Comparing a linear fit with a second-degree and fifth-degree polynomial



- The quadratic polynomial fits the data the best suggesting the model

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

- The quadratic polynomial fits the data the best suggesting the model

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

- Although we use a non-linear function of horsepower this is still alinear model because it is linear in its parameters

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where $X_1 = \texttt{horsepower}$ and $X_2 = \texttt{horsepower}^2$

Probability & Statistics for Computer Science

- The quadratic polynomial fits the data the best suggesting the model

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

- Although we use a non-linear function of horsepower this is still alinear model because it is linear in its parameters

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where $X_1 = \texttt{horsepower}$ and $X_2 = \texttt{horsepower}^2$

Exercise

Fit the quadratic model using the `Auto` data

It seems too good to be true!

- We may encounter problems after we fit a linear regression model

It seems too good to be true!

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model
    1. Non-linearity of the X, Y relationship

It seems too good to be true!

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model
  1. Non-linearity of the X, Y relationship
  2. Correlation of the error terms

It seems too good to be true!

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model
    1. Non-linearity of the X, Y relationship
    2. Correlation of the error terms
    3. Non-constant variance of the error terms

It seems too good to be true!

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model
    1. Non-linearity of the X, Y relationship
    2. Correlation of the error terms
    3. Non-constant variance of the error terms
    4. Outliers

It seems too good to be true!

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model
    1. Non-linearity of the X, Y relationship
    2. Correlation of the error terms
    3. Non-constant variance of the error terms
    4. Outliers
    5. High-leverage points

It seems too good to be true!

- We may encounter problems after we fit a linear regression model
- Many of these challenge the underlying assumptions of our model
  1. Non-linearity of the X, Y relationship
  2. Correlation of the error terms
  3. Non-constant variance of the error terms
  4. Outliers
  5. High-leverage points
  6. Collinearity

1. Non-linearity

1. Non-linearity

- Residual plots are useful for identifying non-linearity

1. Non-linearity

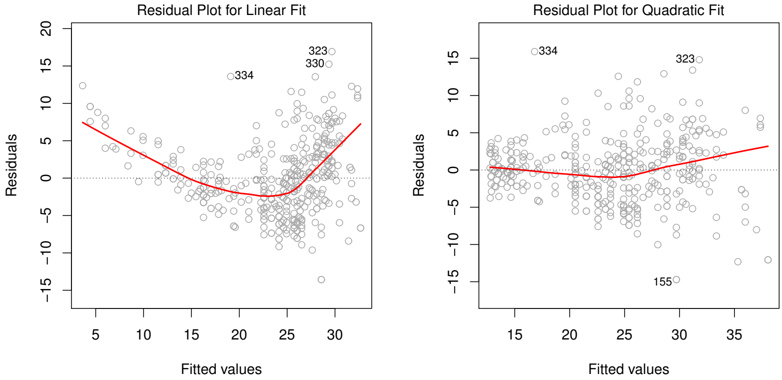
- Residual plots are useful for identifying non-linearity
- Plot the residuals $e_i = y_i - \hat{y}_i$ against x_i for SLR

1. Non-linearity

- Residual plots are useful for identifying non-linearity
- Plot the residuals $e_i = y_i - \hat{y}_i$ against x_i for SLR
- Plot the residuals against $\hat{y}_i$ for MLR

Probability & Statistics for Computer Science

Resdiual Plots left: linear fit; right: quadratic fit



To remove the non-linearity we can transform the predictors with non-linear transformations such as $\log X, \sqrt{X}, X^2$

2. Correlation of the Error Terms

2. Correlation of the Error Terms

- A key assumption for linear regression is that the errors e_i are uncorrelated

2. Correlation of the Error Terms

- A key assumption for linear regression is that the errors e_i are uncorrelated
- If e_i is positive or negative that provides no information about e_{i+1}

2. Correlation of the Error Terms

- A key assumption for linear regression is that the errors e_i are uncorrelated
- If e_i is positive or negative that provides no information about e_{i+1}
- *If* the errors *are* correlated the standard error will be underestimated

2. Correlation of the Error Terms

- A key assumption for linear regression is that the errors e_i are uncorrelated
- If e_i is positive or negative that provides no information about e_{i+1}
- *If* the errors *are* correlated the standard error will be underestimated
- Correlated error terms occur frequently in time series data

2. Correlation of the Error Terms

- A key assumption for linear regression is that the errors e_i are uncorrelated
- If e_i is positive or negative that provides no information about e_{i+1}
- *If* the errors *are* correlated the standard error will be underestimated
- Correlated error terms occur frequently in time series data
- Often observations in adjacent time periods will be correlated

3. Non-constant Variance of the Error Terms

3. Non-constant Variance of the Error Terms

- Another key assumption of linear regression is the constant variance of the error terms

3. Non-constant Variance of the Error Terms

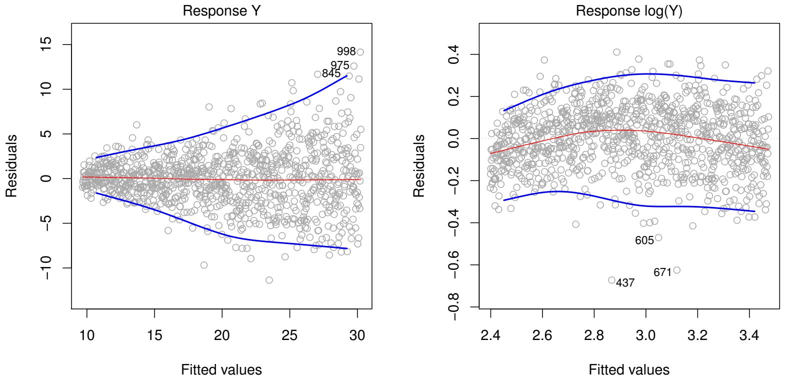
- Another key assumption of linear regression is the constant variance of the error terms
- Sometimes the variance of the error terms is not constant known as *heteroscedasticity*

3. Non-constant Variance of the Error Terms

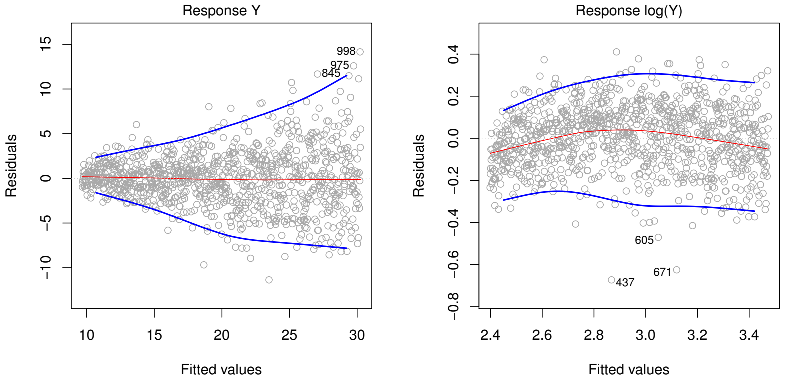
- Another key assumption of linear regression is the constant variance of the error terms
- Sometimes the variance of the error terms is not constant known as *heteroscedasticity*
- If the error terms have non-constant variance the residual plot will show a funnel shape

Probability & Statistics for Computer Science

Residual plots for the original data showing heteroscedasticity (left) and for the transformed data (right)



Residual plots for the original data showing heteroscedasticity (left) and for the transformed data (right)



Transforming Y using $\log Y$ (as in the left plot) or $\sqrt{Y}$ generally reduces heteroscedasticity

4. Outliers

4. Outliers

- An outlier is a point for which y_i is far from $\hat{y}_i$

4. Outliers

- An outlier is a point for which y_i is far from $\hat{y}_i$
- True outliers typically result from some explainable cause like the incorrect recording of data during collection

4. Outliers

- An outlier is a point for which y_i is far from $\hat{y}_i$
- True outliers typically result from some explainable cause like the incorrect recording of data during collection
- If the outlier is due to a data collection error you could eliminate it, but verify it before you do

4. Outliers

- An outlier is a point for which y_i is far from $\hat{y}_i$
- True outliers typically result from some explainable cause like the incorrect recording of data during collection
- If the outlier is due to a data collection error you could eliminate it, but verify it before you do
- An outlier may not affect the fitted regression much but can mess up RSE, p-values, R^2

4. Outliers

- An outlier is a point for which y_i is far from $\hat{y}_i$
- True outliers typically result from some explainable cause like the incorrect recording of data during collection
- If the outlier is due to a data collection error you could eliminate it, but verify it before you do
- An outlier may not affect the fitted regression much but can mess up RSE, p-values, R^2
- Residual plots help identify outliers, we plot Studentized residuals, $e_i/SE(e_i)$, against fitted values

4. Outliers

- An outlier is a point for which y_i is far from $\hat{y}_i$
- True outliers typically result from some explainable cause like the incorrect recording of data during collection
- If the outlier is due to a data collection error you could eliminate it, but verify it before you do
- An outlier may not affect the fitted regression much but can mess up RSE, p-values, R^2
- Residual plots help identify outliers, we plot Studentized residuals, $e_i/SE(e_i)$, against fitted values
- Any data point for which $\left|\frac{e_i}{SE(e_i)}\right| > 3$ is a possible outlier

5. High-leverage Points

5. High-leverage Points

- A high leverage point is an observation y_i that has an unusual value for x_i

5. High-leverage Points

- A high leverage point is an observation y_i that has an unusual value for x_i
- Removing a high-leverage point has a bigger impact on the model than removing an outlier

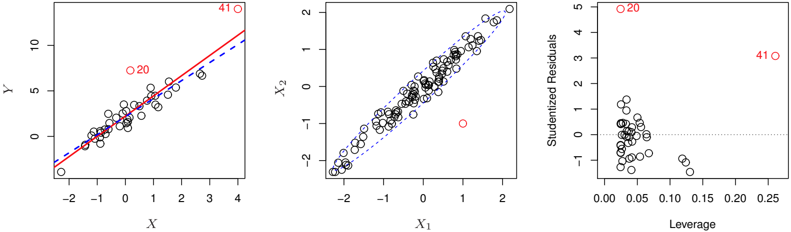
5. High-leverage Points

- A high leverage point is an observation y_i that has an unusual value for x_i
- Removing a high-leverage point has a bigger impact on the model than removing an outlier
- The leverage statistic for SLR is

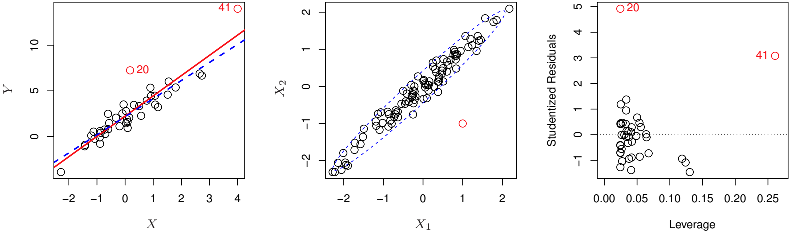
$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \overline{x})^2}$$

which extends to MLR but for thr latter we'll let R calculate it

Probability & Statistics for Computer Science

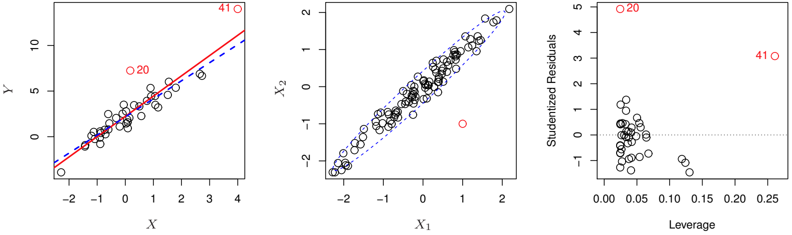


Probability & Statistics for Computer Science



- Point 41 is an outlier *and* a high leverage point that can greatly influence our regression

Probability & Statistics for Computer Science



- Point 41 is an outlier *and* a high leverage point that can greatly influence our regression
- Point 20 is an outlier but has very low leverage thus has little effect on the regression

6. Collinearity

6. Collinearity

- If two or more predictor variables are closely related they are collinear

6. Collinearity

- If two or more predictor variables are closely related they are collinear
- Using the `Credit` data Figure 3.14 (next slide) illustrates collinearity

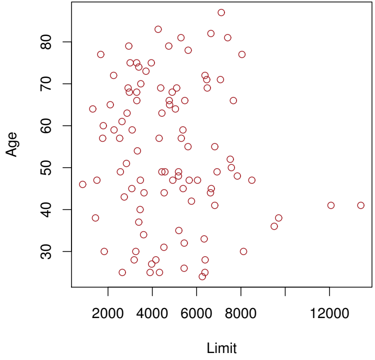
6. Collinearity

- If two or more predictor variables are closely related they are collinear
- Using the **`Credit`** data Figure 3.14 (next slide) illustrates collinearity
- In the left panel **`age`** and **`limit`** have no relationship but on the right it's clear that **`rating`** and **`limit`** are highly correlated

6. Collinearity

- If two or more predictor variables are closely related they are collinear
- Using the `Credit` data Figure 3.14 (next slide) illustrates collinearity
- In the left panel `age` and `limit` have no relationship but on the right it's clear that `rating` and `limit` are highly correlated
- Since they are so strongly correlated it's difficult to determine how each relates to the response

The variables on the left are uncorrelated but those on the right are highly correlated



- Collinearity reduces the accuracy of the $\hat{\beta}$'s

- Collinearity reduces the accuracy of the $\hat{\beta}$'s
- Collinearity reduces the power of the hypothesis test - we may fail to reject $H_0 : \quad \beta_j = 0$ when we should

- Collinearity reduces the accuracy of the $\hat{\beta}$'s
- Collinearity reduces the power of the hypothesis test - we may fail to reject $H_0 : \quad \beta_j = 0$ when we should
- Collinearity causes one variable to be insignificant when another, correlated variable is in the model

Variance Inflation Factor

- We could look at a correlation matrix to find collinearity but this isn't the most efficient way, and it won't show multicollinearity – when three or more variables are correlated

- We could look at a correlation matrix to find collinearity but this isn't the most efficient way, and it won't show multicollinearity – when three or more variables are correlated
- The Variance Inflation Factor (VIF) is a better way

- We could look at a correlation matrix to find collinearity but this isn't the most efficient way, and it won't show multicollinearity – when three or more variables are correlated
- The Variance Inflation Factor (VIF) is a better way

Definition

The VIF is the ratio of the variance of $\hat{\beta}_j$ in the full model to the variance of $\hat{\beta}_j$ fit on its own

- We could look at a correlation matrix to find collinearity but this isn't the most efficient way, and it won't show multicollinearity – when three or more variables are correlated
- The Variance Inflation Factor (VIF) is a better way

Definition

The VIF is the ratio of the variance of $\hat{\beta}_j$ in the full model to the variance of $\hat{\beta}_j$ fit on its own

- $VIF \geq 1$ and $VIF > 5$ or 10 indicates multicollinearity

- We could look at a correlation matrix to find collinearity but this isn't the most efficient way, and it won't show multicollinearity – when three or more variables are correlated
- The Variance Inflation Factor (VIF) is a better way

Definition

The VIF is the ratio of the variance of $\hat{\beta}_j$ in the full model to the variance of $\hat{\beta}_j$ fit on its own

- $VIF \geq 1$ and $VIF > 5$ or 10 indicates multicollinearity

VIF for $\hat{\beta}_j$ is given by

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

VIF for $\hat{\beta}_j$ is given by

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

$R^2_{X_j|X_{-j}}$ is the R^2 from the regression of X_j onto all the other predictor variables

VIF for $\hat{\beta}_j$ is given by

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

$R^2_{X_j|X_{-j}}$ is the R^2 from the regression of X_j onto all the other predictor variables

The closer $R^2_{X_j|X_{-j}}$ is to one, the larger VIF gets, indicating the presence of multicollinearity

Probability & Statistics for Computer Science

R output for a multiple linear regression model

```
Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

References

Miller, I. and Miller, M. (2014).
John E. Freund's Mathematical Statistics with Applications.
Pearson, 8th edition.