

## Day 4: Linear Regression

ME314: Introduction to Data Science and Machine Learning

LSE Methods Summer Programme 2021

24 June 2021

# Day 4 Outline

## Simple linear regression

- Estimation of the parameters

- Confidence intervals

- Hypothesis testing

- Assessing overall accuracy of the model

- Multiple Linear Regression

- Interpretation

- Model fit

## Qualitative predictors

- Qualitative predictors in regression models

- Interactions

- Non-linear effects

## **Simple linear regression**

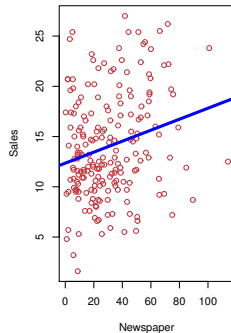
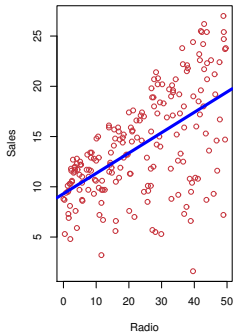
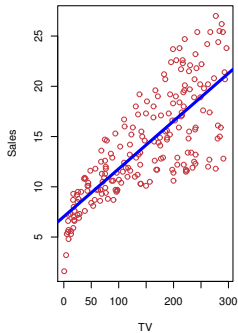
- ▶ Linear regression is a simple approach to supervised learning. It assumes that the dependence of  $Y$  on  $X_1, X_2, \dots, X_p$  is linear.
- ▶ True regression functions are never linear!
- ▶ Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

# Linear regression for the advertising data

Consider the advertising data. Questions we might ask:

- ▶ Is there a relationship between advertising budget and sales?
- ▶ How strong is the relationship between advertising budget and sales?
- ▶ Which media contribute to sales?
- ▶ How accurately can we predict future sales?
- ▶ Is the relationship linear?
- ▶ Is there synergy among the advertising media?

# Advertising data



# Simple linear regression using a single predictor $X$

- ▶ We assume a model

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

where  $\beta_0$  and  $\beta_1$  are two unknown constants that represent the **intercept** and **slope**, also known as **coefficients** or **parameters**, and  $\epsilon$  is the error term.

- ▶ 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.

# Estimation of the parameters by least squares

- ▶ Let  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  be the prediction for  $Y$  based on the  $i$ th value of  $X$ . Then  $e_i = y_i - \hat{y}_i$  represents the  $i$ th **residual**.
- ▶ We define the **residual sum of squares** (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$



# Estimation of the parameters by least squares

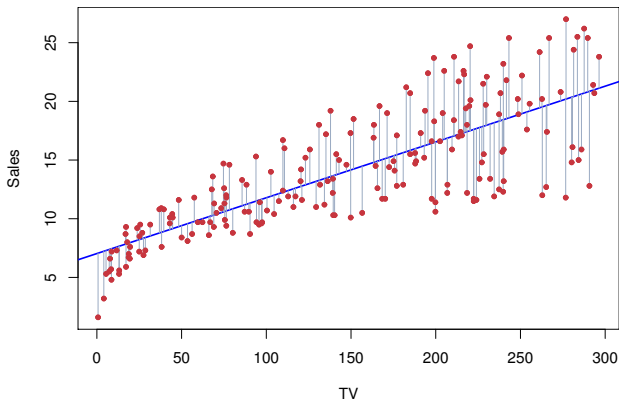
- ▶ The least squares approach chooses  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to minimize the RSS. The minimizing values can be shown to be

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

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

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

## Example: advertising data



The least squares fit for the regression of **sales** on **TV**. The fit is found by minimizing the sum of squared residuals. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

# Assessing the Accuracy of the Coefficient Estimates

- ▶ The standard error of an estimator reflects how it varies under repeated sampling. We have

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

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

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

- ▶ 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

$$\hat{\beta}_1 \pm 2 \times \text{SE}(\hat{\beta}_1).$$

# Confidence Intervals

That is, there is approximately a 95% chance that the interval

$$\left[ \hat{\beta}_1 - 2 \times \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \times \text{SE}(\hat{\beta}_1) \right]$$

will contain the true value of  $\beta_1$  (under a scenario where we got repeated samples like the present sample).

# Hypothesis testing

- ▶ Standard errors can also be used to perform **hypothesis tests** on the coefficients. The most common hypothesis test involves testing the **null hypothesis** of  
 $H_0$ : There is no relationship between  $X$  and  $Y$  versus the **alternative hypothesis**.  
 $H_A$ : There is some relationship between  $X$  and  $Y$ .
- ▶ Mathematically, this corresponds to testing versus

$$H_0 : \beta_1 = 0$$

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  $Y$ .

# Hypothesis testing

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

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)},$$

- ▶ 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**.

# Assessing the Overall Accuracy of the Model

- ▶ We compute the **Residual Standard Error**

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

where the **residual sum-of-squares** is  $\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .

- ▶ **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 linear regression setting that  $R^2 = r^2$ , where  $r$  is the correlation between  $X$  and  $Y$ :

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

## Results for the advertising data

```
advertising <- read.csv("https://www.statlearning.com/s/Advertisi
names(advertising)

## [1] "X"          "TV"          "radio"       "newspaper"  "sales"

simple.regression <- lm(advertising$sales ~ advertising$TV)
```



## Results for the advertising data

```
summary(simple.regression)

##
## Call:
## lm(formula = advertising$sales ~ advertising$TV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.032594   0.457843   15.36  <2e-16 ***
## advertising$TV 0.047537   0.002691   17.67  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

# 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  $\beta_j$  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

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

# 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 a  $\beta_j$  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.

# The woes of (interpreting) regression coefficients

## “Data Analysis and Regression” Mosteller and Tukey 1977

- ▶ a regression coefficient  $\beta_j$  estimates the expected change in  $Y$  per unit change in  $X_j$ , **with all other predictors held fixed**. But predictors usually change together!
- ▶ Example:  $Y$  total amount of change in your pocket;  $X_1$  = number of coins;  $X_2$  = number of pennies, nickels and dimes. By itself, regression coefficient of  $Y$  on  $X_2$  will be  $> 0$ . But how about with  $X_1$  in model?
- ▶  $Y$  = number of tackles by a rugby player in a season;  $W$  and  $H$  are his weight and height. Fitted regression model is  $\hat{Y} = \beta_0 + .50W - .10H$ . How do we interpret  $\hat{\beta}_2 < 0$ ?

## Two quotes by famous Statisticians

- ▶ “Essentially, all models are wrong, but some are useful” George Box
- ▶ “The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively” Fred Mosteller and John Tukey, paraphrasing George Box

# Estimation and Prediction for Multiple Regression

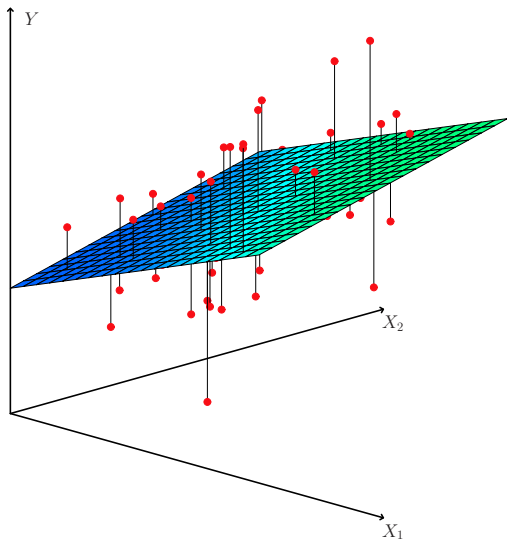
- ▶ Given estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ , we can make predictions using the formula

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

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

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2.$$

This is done using standard statistical software. The values  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  that minimize RSS are the multiple least squares regression coefficient estimates.



## Results for the advertising data

```
multiple.regression <-  
  lm(advertising$sales ~ advertising$TV +  
      advertising$radio + advertising$newspaper)  
cor(advertising[, -1])
```

##		TV	radio	newspaper	sales
## TV		1.00000000	0.05480866	0.05664787	0.7822244
## radio		0.05480866	1.00000000	0.35410375	0.5762226
## newspaper		0.05664787	0.35410375	1.00000000	0.2282990
## sales		0.78222442	0.57622257	0.22829903	1.0000000



## Results for the advertising data

```
summary(multiple.regression)

##
## Call:
## lm(formula = advertising$sales ~ advertising$TV + advertising
##      advertising$newspaper)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.938889    0.311908   9.422  <2e-16 *
## advertising$TV  0.045765    0.001395  32.809  <2e-16 *
## advertising$radio  0.188530    0.008611  21.893  <2e-16 *
## advertising$newspaper -0.001037    0.005871  -0.177    0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Some important questions

1. Is at least one of the predictors  $X_1, X_2, \dots, X_p$  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 can use the F-statistic

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

## 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 there 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.

## Model selection - continued

- ▶ Later we discuss more systematic criteria for choosing an “optimal” member in the path of models produced by forward or backward stepwise selection.
- ▶ These include Mallow's  $C_p$ , Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted  $R^2$  and Cross-validation (CV).

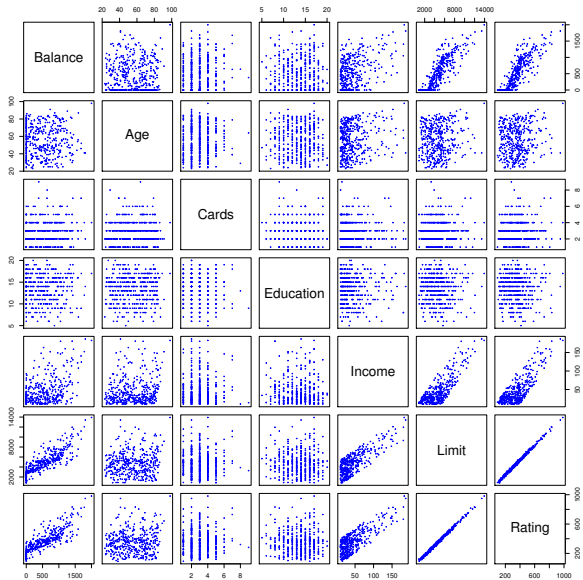
## **Qualitative predictors**



# Other Considerations in the Regression Model

## 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**.
- ▶ See for example the scatterplot matrix of the credit card data in the next slide.
- ▶ 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).



## Qualitative Predictors – continued

- ▶ Example: 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?

## Credit card data

```
credit <- read.csv("https://www.statlearning.com/s/Credit.csv")
names(credit)

##  [1] "X"          "Income"     "Limit"      "Rating"     "Cards"
##  [7] "Education" "Gender"     "Student"    "Married"    "Ethnici

gender.regression <- lm(credit$Balance ~ credit$Gender)
```

## Results for gender model

```
summary(gender.regression)

##
## Call:
## lm(formula = credit$Balance ~ credit$Gender)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -529.54 -455.35  -60.17   334.71 1489.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      529.54       31.99  16.554  <2e-16 ***
## credit$GenderMale -19.73       46.05  -0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611, Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

## Qualitative predictors with more than two levels

- ▶ 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 could be

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

## Qualitative predictors with more than two levels

- ▶ Then both of these variables can be used in the regression equation, in order to obtain the model

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

- ▶ 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**.

## Credit card data

```
ethnicity.regression <- lm(credit$Balance ~ credit$Ethnicity)
summary(ethnicity.regression)

##
## Call:
## lm(formula = credit$Balance ~ credit$Ethnicity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -531.00 -457.08  -63.25   339.25 1480.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      531.00      46.32   11.464  <2e-16
## credit$EthnicityAsian    -18.69      65.02   -0.287    0.77
## credit$EthnicityCaucasian -12.50      56.68   -0.221    0.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
```



# Extensions of the Linear Model

Removing the additive assumption: **interactions** and **nonlinearity**

**Interactions:**

- ▶ In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.
- ▶ For example, the linear model

$$\widehat{sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper$$

states that the average effect on sales of a one-unit increase in TV is always  $\beta_1$ , regardless of the amount spent on radio.

## Interactions – continued

- ▶ But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- ▶ In this situation, given a fixed budget of \$100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- ▶ In marketing, this is known as a **synergy** effect, and in statistics it is referred to as an **interaction** effect.

## Modelling interactions – Advertising data

Model takes the form

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \epsilon$$

$$= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + \epsilon$$

## Modelling interactions – Advertising data

```
interaction.model <- lm(advertising$sales ~ advertising$TV*advertising$radio)
summary(interaction.model)
```

##

## Call:

## lm(formula = advertising\$sales ~ advertising\$TV \* advertising\$radio, data = advertising, weights = NULL)

## Residuals:

	Min	1Q	Median	3Q	Max
##	-6.3366	-0.4028	0.1831	0.5948	1.5246

##

## Coefficients:

	Estimate	Std. Error	t value
## (Intercept)	6.750e+00	2.479e-01	27.233
## advertising\$TV	1.910e-02	1.504e-03	12.699
## advertising\$radio	2.886e-02	8.905e-03	3.241
## advertising\$TV:advertising\$radio	1.086e-03	5.242e-05	20.727

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residuals: [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000

# Interpretation

- ▶ The results in this estimation suggests that interactions are important (statistically at least – it may be unimportant substantively)
- ▶ The p-value for the interaction term  $TV \times radio$  is extremely low, indicating that there is strong evidence for  $H_A : \beta_3 \neq 0$ .
- ▶ The  $R^2$  for the interaction model is 96.8%, compared to only 89.7% for the model that predicts *sales* using *TV* and *radio* without an interaction term.

## Interpretation – continued

- ▶ This means that  $(96.8 - 89.7)/(100 - 89.7) = 69\%$  of the variability in sales that remains after fitting the additive model has been explained by the interaction term.
- ▶ The coefficient estimates in the table suggest that an increase in TV advertising of \$1,000 is associated with increased sales of

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio units}.$$

- ▶ An increase in radio advertising of \$1,000 will be associated with an increase in sales of

$$(\hat{\beta}_2 + \hat{\beta}_3 \times TV) \times 1000 = 29 + 1.1 \times TV \text{ units}.$$

# Hierarchy

- ▶ Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, *TV* and *radio*) do not.
- ▶ The hierarchy principle: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

# Hierarchy

- ▶ The rationale for this principle is that interactions are hard to interpret in a model without main effects – their meaning is changed.
- ▶ Specifically, the interaction terms also contain main effects, if the model has no main effect terms.



# Interactions between qualitative and quantitative variables

- ▶ Consider the *Credit* dataset, and suppose that we wish to predict *balance* using *income* (quantitative) and *student* (qualitative).
- ▶ Without an interaction term, the model takes the form

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

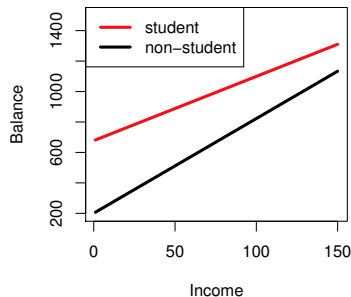
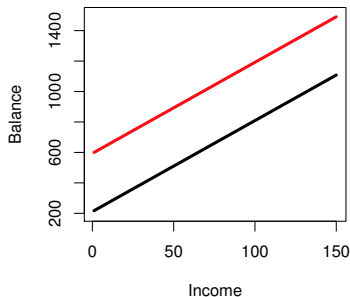
$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases}$$

- With interactions, it takes the form

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 + \beta_3 \times income_i & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

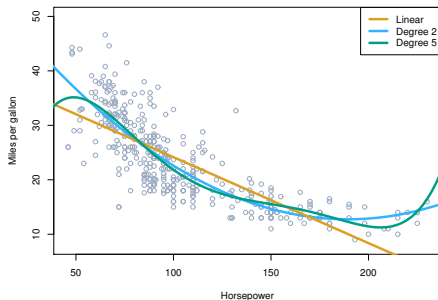
$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income_i & \text{if } i\text{th person is a student} \\ \beta_0 + \beta_1 \times income_i & \text{if } i\text{th person is not a student} \end{cases}$$

# Credit data



- ▶ For the *Credit* data, the least squares lines are shown for prediction of balance from income for students and non-students.
- ▶ Left: no interaction between income and student.
- ▶ Right: with an interaction term between income and student.

# Non-linear effects of predictors



- ▶ Polynomial regression on *Auto* data
- ▶ The figure suggests that

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

may provide a better fit.

```

library(ISLR)
auto.model <- lm(Auto$mpg ~ Auto$horsepower + I(Auto$horsepower^2)
summary(auto.model)

##
## Call:
## lm(formula = Auto$mpg ~ Auto$horsepower + I(Auto$horsepower^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    56.9000997   1.8004268    31.60  <2e-16 *
## Auto$horsepower -0.4661896   0.0311246   -14.98  <2e-16 *
## I(Auto$horsepower^2)  0.0012305   0.0001221    10.08  <2e-16 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686

```

# What we did not cover

- ▶ Correlation of the error-terms.
- ▶ Non-constant variance of error terms.
- ▶ Outliers.
- ▶ High leverage points.
- ▶ Collinearity.

See text Section 3.3.3

# Generalizations of the Linear Model

In much of the rest of this course, we discuss methods that expand the scope of linear models and how they are fit:

- ▶ **Classification problems:** logistic regression, support vector machines
- ▶ **Non-linearity:** kernel smoothing, splines and generalized additive models; nearest neighbor methods.
- ▶ **Interactions:** Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- ▶ **Regularized fitting:** Ridge regression and lasso