Topic # 02: Data Analytics

Regression: fit, interpretation and prediction

Instructor: Prof. Arnab Bisi, Ph.D.

Johns Hopkins Carey Business School

Regression – Review, Methods, Extensions

Session 2:

Agenda

- Part 1: Regression basics - review
- Break
- Part 2: Regression models in **R**
- Part 3: Working with code

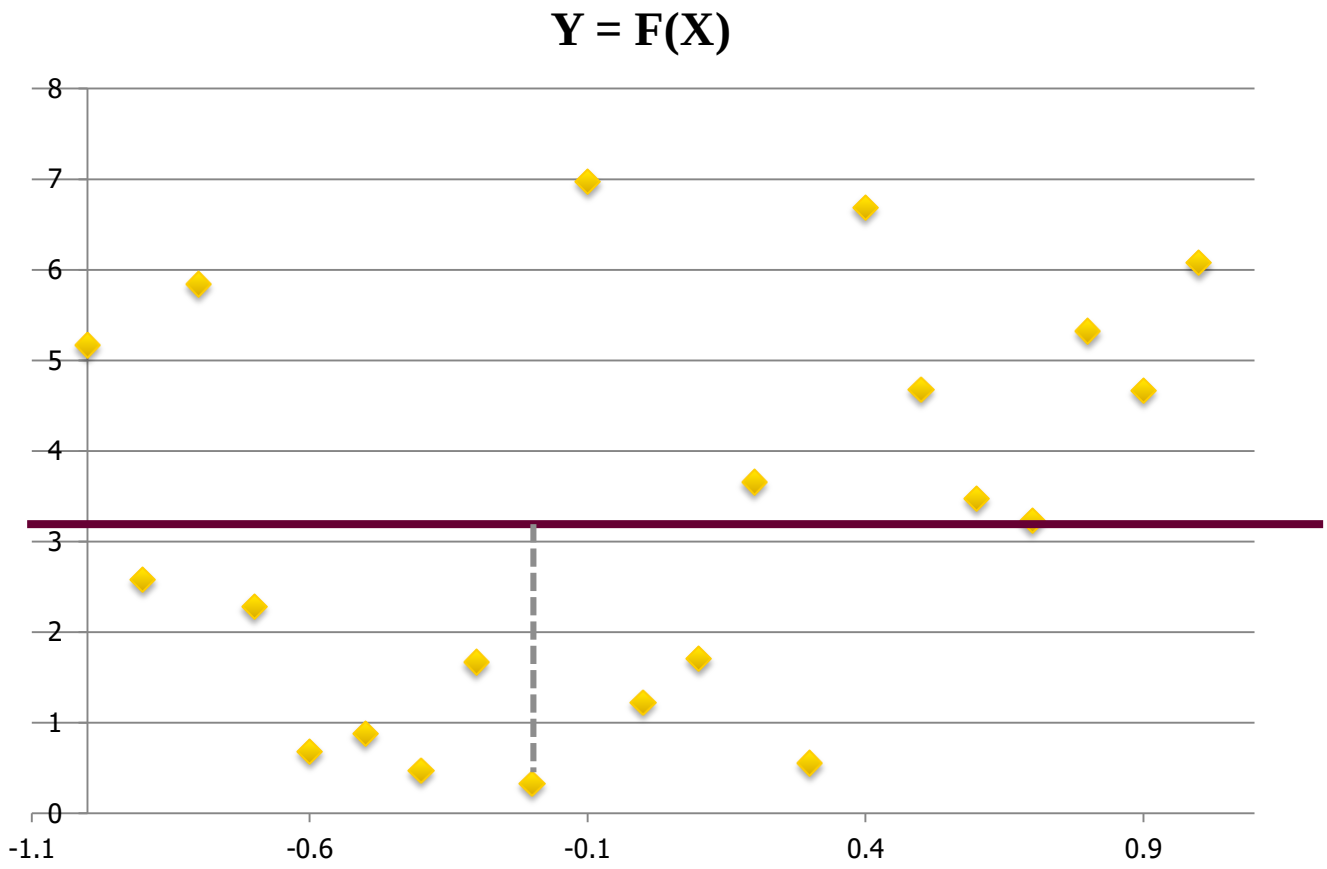
JOHNS HOPKINS
CAREY BUSINESS SCHOOL

What Is Regression?

- Regression modeling is one of the most useful statistical techniques
- Explanatory variables vs. Response variables
- Allows “what-if” questions
- Linear regression: a fundamental starting point for all regression methods

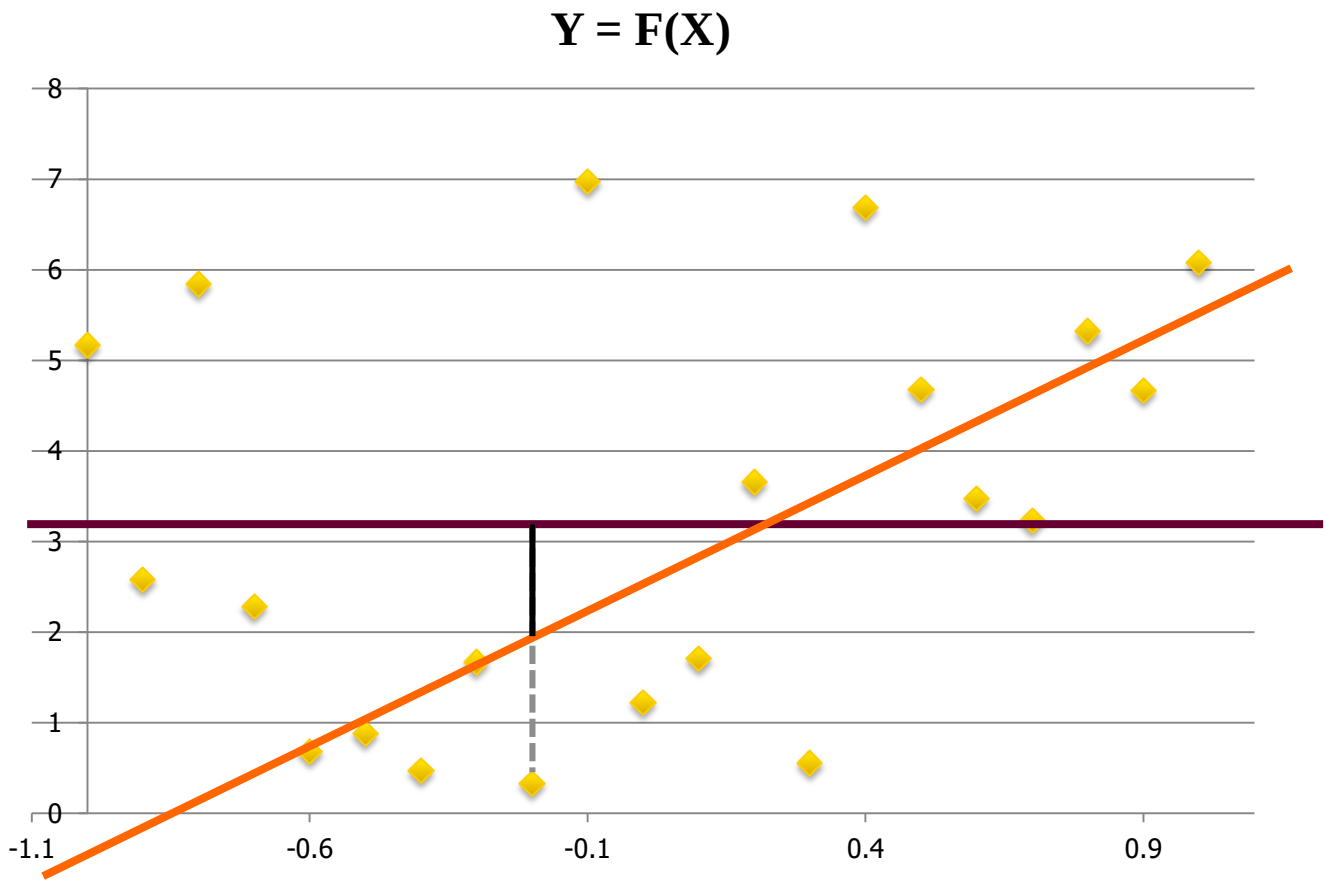
Consider a Predictor and a Response

- For each point *i* we can calculate $(y_i - \bar{y})$^2
  - Some texts call this deviance or squared deviance
- If we total these values we have the Total Sum of Squares = TSS



Consider an Equation for a Line

- B_0 = -0.5 is the intercept
- B_1 = 3 is the slope
- For each point *i* we can calculate $(y_i - y)$^2
  - This error is left over or "Residual" after fitting the line
- If we total these values we have the Residual Sum of Squares = RSS



Estimate the Coefficients

- Denote n observation pairs as follows

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

- Let $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ be the prediction for Y based on the i-th value of X
- Let $e_i = y_i - \widehat{y}_i$ represent the i-th residual, (positive or negative)
- We define the residual sum of squares (RSS) as

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2.$$

or equivalently as

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

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Estimate the Coefficients

- Minimize residual sum of squares (RSS) w.r.t. $\widehat{\beta}_0, \widehat{\beta}_1$, i.e.,

$$\min_{\widehat{\beta}_0, \widehat{\beta}_1} \; (y_1 - \widehat{\beta}_0 - \widehat{\beta}_1 x_1)^2 + \cdots + (y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n)^2.$$

- Consider *first-order-conditions*

$$\sum_{i=1}^{n} (y_i - (\widehat{\beta}_0 - \widehat{\beta}_1 x_i)) = 0 \;\; \text{and} \;\; \sum_{i=1}^{n} x_i (y_i - (\widehat{\beta}_0 - \widehat{\beta}_1 x_i)) = 0$$

- By the least squares approach, (after some algebra)

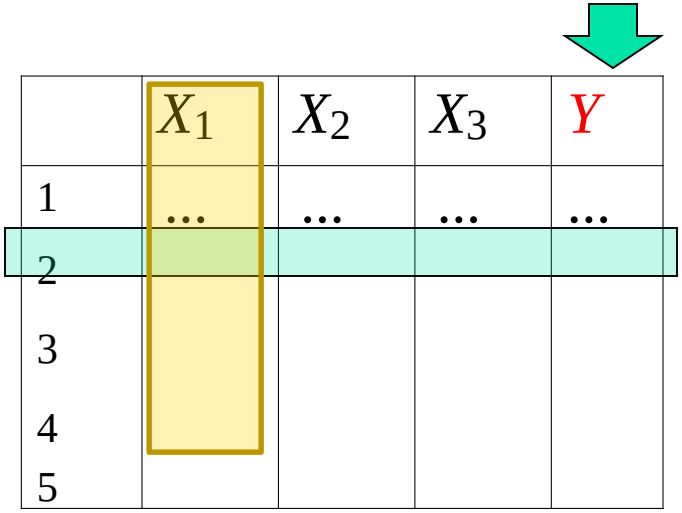
$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}, \end{aligned}$$

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

Linear Regression

- Linear regression is an approach to model the relationship between a scalar dependent variable y and one or more explanatory variables denoted as $\mathbf{X}$
- The case with one explanatory variable is called simple linear regression
- For more than one explanatory variable, it is called multiple linear regression

What Data Looks Likein **R**



- Data/points/instances/examples/samples/records: rows
- Input variables/Independent Variables/features/attributes/dimensions/covariates/predictors/regressors/factors: columns
- Output variable/ outcome/response/label/dependent variable: special column to be predicted

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Main Types of Columns

	X_1	X_2	X_3	Y
1	…	…	…	…
2				
3				
4				

- Continuous: quantitative, a number like weight or length; numeric
- Discrete: qualitative, a symbol like ‘cat” or “dog”, “0”, or “1”, small, medium, large, etc.; factor
- In **R** this entire table is called a Data Frame

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Main Goal of Learning: Prediction

The setup:

Obtain some kind of model based on observations, or training data $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$, through a process called learning (or estimation).

Use that model to predict something about data you haven't seen before, but that comes from the same distribution as the training data, called test data.

What we really care about is the quality of our predictions!!!

Example: Advertising Data Set

Sales (thousand units), advertising budget on TV ($), Radio ($), Newspaper ($)

200 records

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
...				

JOHNS HOPKINS CAREY BUSINESS SCHOOL

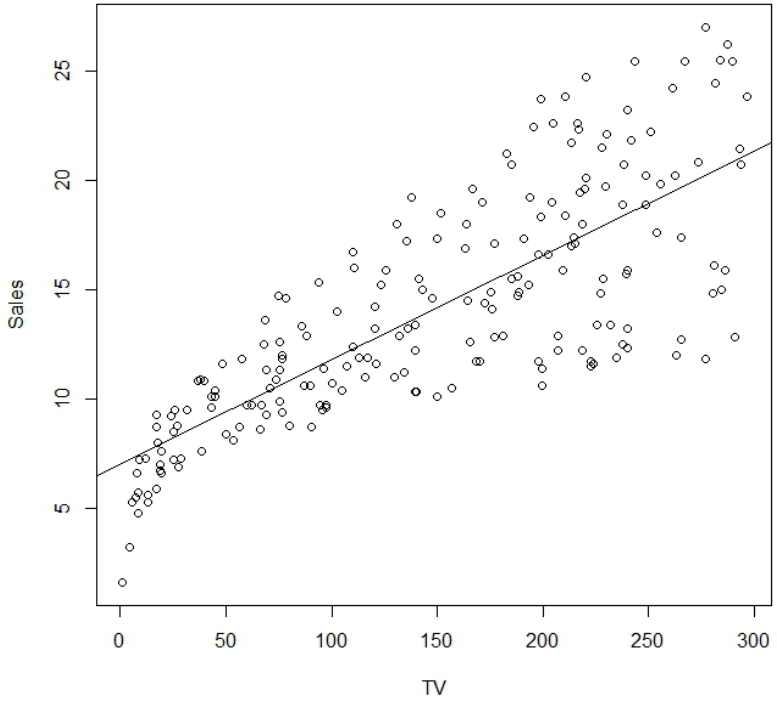
Questions about Advertising Data: Discussion

- ✓ Is there a relationship between advertising budget and sales?
- ✓ How strong is the relationship between advertising budget and sales?
- ✓ Which media contributes most to sales?
- ✓ How accurately can we estimate the effect of each media on sales?
- ✓ How accurately can we predict future sales?
- ✓ Is the relationship linear?

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Advertising Data: Plots

- > ad=read.csv("Advertising.csv")
- > attach(ad)
- > plot(TV,Sales)
- > plot(Sales~Radio)
- > plot(Sales~Newspaper, data=ad)



JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Steps in **R**

- Read data into **R**
  - CSV file from Blackboard
- Plot variables of interest
- Create linear model
  - model.name=lm(y~ x1 + x2 + x3, data=data.frame)
- Consider model output
  - summary(model.name)
  - plot(y~x)
  - abline(model.name)

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Exercise

- Use Advertising data set
- Study the relationship between Sales and Newspaper
  - Produce scatter plot
  - Run linear regression
  - Add regression line
- 10 mins

Example: Data Analytics with R

- Task: Predict sales for advertising budget (100, 50, 25)
- Regression using **R**

Data Analytics: easy and simple

```
predict(lm(Sales ~ TV+Radio+Newspaper, data=ad),
data.frame(TV=100, Radio=50, Newspaper=25),
interval="confidence")
```

- Prediction $16.92 with 95% confidence interval (16.36, 17.47)

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Simple Linear Regression

- Straightforward approach for predicting a quantitative response Y on the basis of a single (univariate) predictor variable X
- It assumes that there is an approximately linear relationship between X and Y
- Mathematically, we can write this linear relationship as

  $Y \approx \beta_0 + \beta_1 X$ or $Y = \beta_0 + \beta_1 X + E$
- Linear regression includes several assumptions
  - The error term has a mean of 0
  - The error terms have a constant variance
  - The errors are uncorrelated
    - If we assume a linear relationship AND the errors have a mean of 0 AND the variance is constant, we have support for the assumption that the relationship is indeed linear

If these assumptions do not hold, this is the WRONG model!!!

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Simple Linear Regression

- Unknown constants: β_0 represents the intercept; β_1 represents slope.
- Use training data to produce estimates of β_0 and β_1
- Then use estimates to make predictions
- Regression models give results that are:
- B – Best
- L – Linear
- U – Unbiased
- E – Estimators

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Simple Linear Regression

- True values of β_0 and β_1 unknown
- Use training data to produce estimates of β_0 and β_1

$$(1) \qquad \hat{Y}_i \approx \hat{b}_0 + \hat{b}_1 x_i$$

- This prediction is the mean of a distribution
  - That is why equation (1) is an approximation

$$(2) \qquad \hat{Y}_i = \hat{b}_0 + \hat{b}_1 x_i + E$$

- **Additional assumption**: error terms are normally distributed
  - This additional assumption greatly simplifies hypothesis testing

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Hypothesis Tests

Null hypothesis

H_0 : there is no linear relationship between X and Y

Alternative hypothesis

H_a : there is some linear relationship between X and Y

Mathematically,

$$H_0 : \beta_1 = 0,$$

vs.

$$H_a : \beta_1 \neq 0.$$

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Hypothesis Tests

- Regression models actually predict the mean of a normal distribution with constant variance

- We compute a t-test statistic given by

$$t = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)}$$

  - It has a t-distribution with n – 2 degrees of freedom
  - This allows the calculation of *p*-values

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Hypothesis Tests

- p-value: 2.the probability of observing any value larger than |t|, assuming $\beta_1 = 0$.
- If the p-value is small enough we reject the null hypothesis and infer that there is a linear relationship between the predictor and the response variable.
  - Typical cutoff values are 5% or 1%

Accuracy of the Model

The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.

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

where RSS formula (*residual sum of squares*) is

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

JOHNS HOPKINS CAREY BUSINESS SCHOOL

R^2: the Coefficient of Determination

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

- R^2 measures the proportion of variability in Y that has been explained by the model that uses **X**
  - If the model is a perfect fit, RSS will be 0
    - When RSS = 0, $R^2 = 1$
    - By using X we have eliminated all of the errors
    - This is as good as can be done
  - If the model is useless, the new model is no better than the null model
    - RSS will be no lower than TSS
    - If RSS = TSS , $R^2 = 0$
    - Can be because the model is wrong
    - Can occur when the inherent error is very high

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Adding Terms to a Regression Model

- Use data to produce estimates of β_0 and β_1

$$\hat{Y}_i \approx \hat{b}_0 + \hat{b}_1 x_i$$

- Consider adding a new predictor, z

- We now estimate β_0, β_1 and β_2 to get

$$\hat{Y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + E$$

- Note that the objective function is still the same
  - There is no constraint that prevents us from using the old estimates of β_0 and β_1
  - Since $\hat{b}_2$ can equal any value we like, we would never select a value that makes the objective worse
    - Therefore adding a new variable can NEVER reduce R^2 even if the new variable is meaningless

JOHNS HOPKINS CAREY BUSINESS SCHOOL

R^2 and Adjusted R^2

By adding variables to a model, the *residual sum of squares* can only (RSS) decrease, so the R^2 increases.

The adjusted R^2 for a model with p predictors and $p+1$ estimated coefficients is defined as,

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

This introduces a penalty for the number of estimated coefficients.

While the R^2 can never decrease as more variables are added to the model, the adjusted R^2 with too many unneeded variables can actually decrease.

Confidence Intervals

To compute the standard errors associated with $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we use the following formulas:

$$\begin{aligned} SE(\widehat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right], \\ SE(\widehat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \end{aligned}$$

where $\sigma = Var(\epsilon)$ can be estimated by the *residual standard error* (RSE), i.e.,

$$RSE = \sqrt{RSS/(n-2)}$$

Then, the 95% confidence intervals are

$$\begin{aligned} &[\widehat{\beta}_0 - 2 \cdot SE(\widehat{\beta}_0),\ \widehat{\beta}_0 + 2 \cdot SE(\widehat{\beta}_0)], \\ &[\widehat{\beta}_1 - 2 \cdot SE(\widehat{\beta}_1),\ \widehat{\beta}_1 + 2 \cdot SE(\widehat{\beta}_1)]. \end{aligned}$$

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Prediction Interval

In linear regression,

$$Var(y) = Var(\widehat{y} + \epsilon) = Var(\widehat{y}) + \sigma^2.$$

The 95% prediction interval is $\widehat{y} \pm 2\sqrt{\sigma^2 + SE^2}$.

So, the prediction intervals are ALWAYS Wider than the confidence intervals.

Multiple Linear Regression

More than one predictor.

Multiple linear regression model

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

where X_j represents the j-th predictor and β_j quantifies the association between that variable and the response.

We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.

This is only valid if the predictors are not correlated

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Estimate the Regression Coefficients

- Given estimates, we make prediction using the formula

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

- Denote n observation pairs as follows

$$(\mathbf{x}_1, y_1),\ (\mathbf{x}_2, y_2), \ldots,\ (\mathbf{x}_n, y_n),$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

- Least Squares Approach, minimize RSS

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2. \\ &= \sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + \widehat{\beta} x_{i1} + \widehat{\beta} x_{i2} + \cdots + \widehat{\beta} x_{ip})\right)^2. \end{aligned}$$

Hypothesis: Multiple Linear Regression

Is there a relationship between the response and predictors?

Null hypothesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0.$

Alternative hypothesis

H_a : at least one β_j is non-zero.

This hypothesis test is performed by computing the F-statistic,

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

Large F-statistic provides evidence against the null hypothesis H_0.

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Understanding the Regression Summary

call: this shows how lm() was called

Residuals statistics: Min and Max; first quartile (1Q) and third quartile (3Q); median

Coefficients

The column labeled Estimate contains the estimated regression coefficients as calculated by ordinary least squares.

The column labeled Std. Error is the standard error of the estimated coefficient. The column labeled t value is the t statistic from which the p-value was calculated.

Question: statistically speaking, how likely is it that the true coefficient is zero?

The p-value is a probability. It gauges the likelihood that the coefficient is not significant, so **smaller** is **better**.

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Understanding the Regression Summary

Residual standard error: this reports the standard error of the residuals – that is, the sample standard deviation.

R^2 and adjusted R^2: R^2 is a measure of the model's quality; the adjusted R^2 accounts for the number of variables in your model and so is a more realistic assessment of its effectiveness. **Bigger** is **better**.

F statistic: the F statistic tells you whether the model is significant or insignificant. The model is significant if any of the coefficients are nonzero. Conventionally, a p-value of less than **0.05** indicates that the model is likely significant (one or more β_i are nonzero)

Most people look at the R^2 statistic first. The statistician wisely starts with the F statistic, for if the model is not significant then nothing else matters.

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Extensions of the Linear Model

Remove additional assumption

Standard linear regression model with two variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E.$$

Interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + E.$$

Non-linear Relationships: example

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + E$$

quadratic model, polynomial model (will talk more in the future)

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Questions, Comments?

Let's move to the Code.