Class 2: Linear Regression

MSA 8150: Machine Learning for Analytics

Alireza Aghasi

Institute for Insight, Georgia State University

Important Note About the TA and Homework Submission

- **Teaching Assistant:** Maohua Xie (mxie3@student.gsu.edu)
  - Issues related to the homework grades should be discussed with the TA
  - If an issue was not resolved, then it could be discussed with the instructor during the office hours
- **(IMPORTANT UPDATE):** to be consistent and upon discussion with the TA, **all homework submissions will be through the iCollege (no in-class hard copies)**

A Brief Review of Hypothesis Testing

Some Basic Probability Overview

- For a continuous random variable X we often define a probability density distribution $f_X(x)$ where

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$$

- A random variable X is normally distributed with mean μ and variance σ^2 (denoted as $\mathcal{N}(\mu, \sigma^2)$), when

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Expectation of the weighted sum: if $x_1, \ldots, x_n$ are random variables with mean $\mathbb{E}(x_i) = \mu_i$, then for constants α_i:

$$\mathbb{E}(\alpha_1 x_1 + \ldots \alpha_n x_n) = \alpha_1\mu_1 + \ldots + \alpha_n\mu_n$$

- Variance of the weighted sum: if $x_1, \ldots, x_n$ are **independent** random variables with variance $var(x_i) = \sigma_i^2$, then

$$var(\alpha_1 x_1 + \ldots \alpha_n x_n) = \alpha_1^2\sigma_1^2 + \ldots + \alpha_n^2\sigma_n^2$$

Hypothesis Testing

- We call H_0 as the **null** hypothesis and refer to H_1 as the **alternative** hypothesis
- The hypothesis we want to test is if H_1 is likely to be true
- Often, the equality hypothesis is chosen to be the null hypothesis

Hypothesis Testing

- A hypothesis is often a conjecture about one or more populations
- To prove a hypothesis is true we need to examine all the population which is often not practical, **instead we take a random sample** and probabilistically assess if we have enough evidence to support a hypothesis
- **Example**: All human beings respond well to a specific treatment
  - Instead of testing it on all people on the planet, we look into a fraction of people who are infected by the decease and see if the treatment works
- Since we are focused on a limited sample, we can only state our confidence about the conjecture probabilistically

Hypothesis Testing

- Hypothesis testing is often formulated in terms of two hypotheses
  - H_0: the null hypothesis
  - H_1: the alternate hypothesis
- We want to test if H_1 is likely to be true
- You decide to make a claim about the null hypothesis H_0 and be certain what you claim is true with probability at least $1 - \alpha$
- **Example**: We are 90% confident that this drug works on patients with xxx decease
- So normally you decide on how confident you want to make a claim and determine a value for α

Hypothesis Testing

In hypothesis testing only one of these two cases happens:

- You **reject** H_0 and **accept** H_1 since you have enough evidence in favor of H_1
- You **fail to reject** H_0, since you don not have enough evidence to support H_1

While you see a lot of documents talking about "*accepting* H_0", technically you should use the term *"fail to reject"*

- It might be the case that H_0 is false, but your data is not enough to reject it (does not mean you should accept it)

Example: H_0: Tim is innocent H_1: Tim is guilty

If you have enough to support Tim is guilty, you reject H_0. If you do not have enough to show that Tim is guilty (failure to reject H_0), that does not mean he is innocent (accepting H_0)

Hypothesis Testing: Types of Error

	Reject H_0 (accept H_1)	Do not reject H_0
H_0 true in reality (probability)	Type I error α	Correct decision $1-\alpha$
H_1 true in reality (probability)	Correct decision $1-\beta$	Type II error β

- α is a small number that we determine and is called the significance level (the probability of making type I error)
- We decide on how confident we want to make a claim in favor of H_0 and $1-\alpha$ is our confidence about this
- Normally people take α to be 0.05 or 0.01, giving you 95% or 99% chance of validity in making the argument in support of H_0
- We also have a type II error (calling $1-\beta$ the power of the test), but here we do not want to focus on that

Hypothesis Testing Example

- In the context of our linear regression problem we are interested in hypothesis testing problems on the basis of samples

Example: There is a normal distribution with variance 1 and unknown mean μ. We take 10 independent samples x_i of this distribution as:

`1.8978, 1.7586, 2.3192, 2.3129, 1.1351, 1.9699, 1.8351, 2.6277, 3.0933 , 3.1093`

we add up these numbers and divide it by 10 (taking the sample mean) and observe that

$$\frac{x_1 + \ldots x_{10}}{10} = 2.2059.$$

We get a feeling that probably $\mu = 2$, so we decide to test this hypothesis:

$$H_0 : \mu = 2 \qquad vs \qquad H_1 : \mu \neq 2 \ \text{(two sided test)}.$$

Hypothesis Testing Example

Solution: We generally look into the behavior of the random variable

$$\bar{x} = \frac{x_1 + \ldots + x_{10}}{10}$$

which is a normally distributed random variable with mean μ and variance 0.1 [can you say why?]. As a result, $z = \frac{\bar{x}-\mu}{\sqrt{0.1}}$ is a standard $\mathcal{N}(0,1)$ random variable [can you say why?].

We refer to z as the **test statistic**.

p-value: *is a useful quantity in the analysis of the test and is the probability of obtaining a result equal or "more extreme" than what we have observed, given that the null hypothesis is true. In the case of this example:*

$$\text{p-value} = \mathbb{P}\left(|z| \geq \frac{2.2059 - \mu}{\sqrt{0.1}} \,\middle|\, \mu = 2\right) = \mathbb{P}\left(|z| \geq 0.6511\right)$$
$$= 0.5150.$$

Hypothesis Testing Example

Suppose our significance level is $\alpha = 0.05$.

> If p-value$\leq \alpha$: reject H_0 (accept H_1)
> If p-value$> \alpha$: fail to reject H_0

For our example p-value $= 0.5150 > 0.05$, so we cannot reject the hypothesis that $\mu = 2$.

- If the value of $\bar{x} = 2.2059$ was obtained by taking the sample mean over 100 samples then we had $z = \frac{\bar{x}-\mu}{\sqrt{0.01}}$ and

$$\text{p-value} = \mathbb{P}\left(|z| \geq \frac{2.2059-\mu}{\sqrt{0.01}} \,\middle|\, \mu = 2\right) = \mathbb{P}\left(|z| \geq 2.059\right)$$
$$= 0.0395 < 0.05,$$

  then we were able to reject H_0.
- In other words we are more than 95% confident (accurately, 96.05% confident) that it is not possible to take the sample mean over 100 random numbers of mean $\mu = 2$ and variance 1, and get a value as far from 2 as 2.2059! [Lets try it on Matlab]

Hypothesis Testing Example: Unknown Variance

Suppose in the previous example we did not know σ and instead of working with the standard random normal variable $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ we work with the random variable

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \qquad \text{where :} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Note that s is an unbiased estimate of the standard deviation (since we don't know σ this is the best we can use).

If we want to go through a similar hypothesis test, we need to look into the test statistic $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ which is no more normally distributed. It has a more complicated distribution called *Student's t distribution*.

So for the probabilities needed to calculate the p-value, we need to refer to t-distribution tables instead of the normal distribution tables.

Your Take Away from Hypothesis Testing

- We have some independent samples from a distribution and we guess something about our data
- We form a hypothesis test with some null and alternate hypotheses
- We fix some value for the significance α, meaning that $1 - \alpha$ is how confident we want to be in making our claim
- We form a test statistic (the resulting random variable can have a very complicated distribution)
- We calculate the p-value:
  - If p-value$\leq \alpha$: reject H_0 (accept H_1)
  - If p-value$> \alpha$: fail to reject H_0

Now Lets Start Linear Regression!

Introduction to Linear Regression

- You remember we had an ideal "regression function" $f(\boldsymbol{x})$, which was the actual function behind our data generation and our observations y were in the form

$$y = f(\boldsymbol{x}) + \epsilon$$

- Note that our input vector $\boldsymbol{x}$ contained all the input features, i.e.,

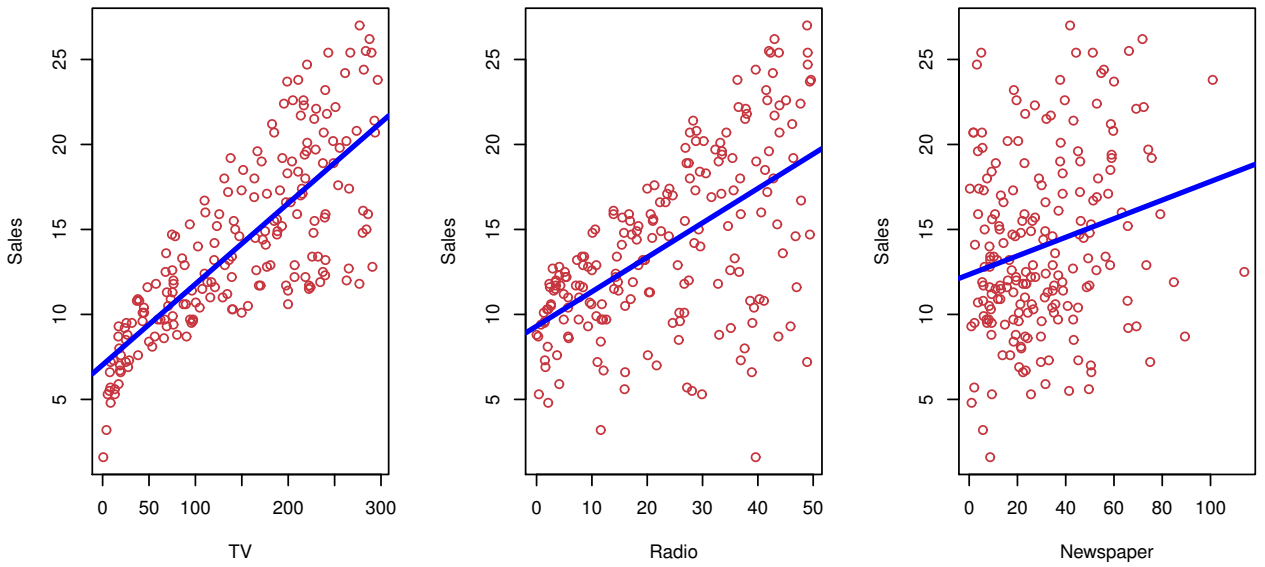
$$\boldsymbol{x} = [x_1, \ldots, x_p]^\top$$

- We had no access to $f(\boldsymbol{x})$, but we wanted to estimate it with some function $\hat{f}$
- In **linear regression** we search for a $\hat{f}$, which takes the following form

$$\hat{f}(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Introduction to Linear Regression

Lets start with a very simple model that only uses one feature



$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \epsilon$$

or more generally,

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where x is the only available feature

More on Simple Linear Regression

- We consider the model

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients and ϵ is the error term.
- We are given samples of the form $(x_1, y_1), \ldots, (x_n, y_n)$, using which we try to fit some values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients
- After this fit we can predict future responses to a test sample x_t, using

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

Determining the Model Coefficients

– Remember that we had samples $(x_1, y_1), \ldots (x_n, y_n)$ and we would like to determine $\hat{f}(x)$ by

$$\min \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

– Using our simple model we would like to decide β_0 and β_1 such that the **Residual Sum of Squares** (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized

Determining the Model Coefficients

– Taking the derivative with respect to β_0 and β_1 and setting it to zero, for $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ we get

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

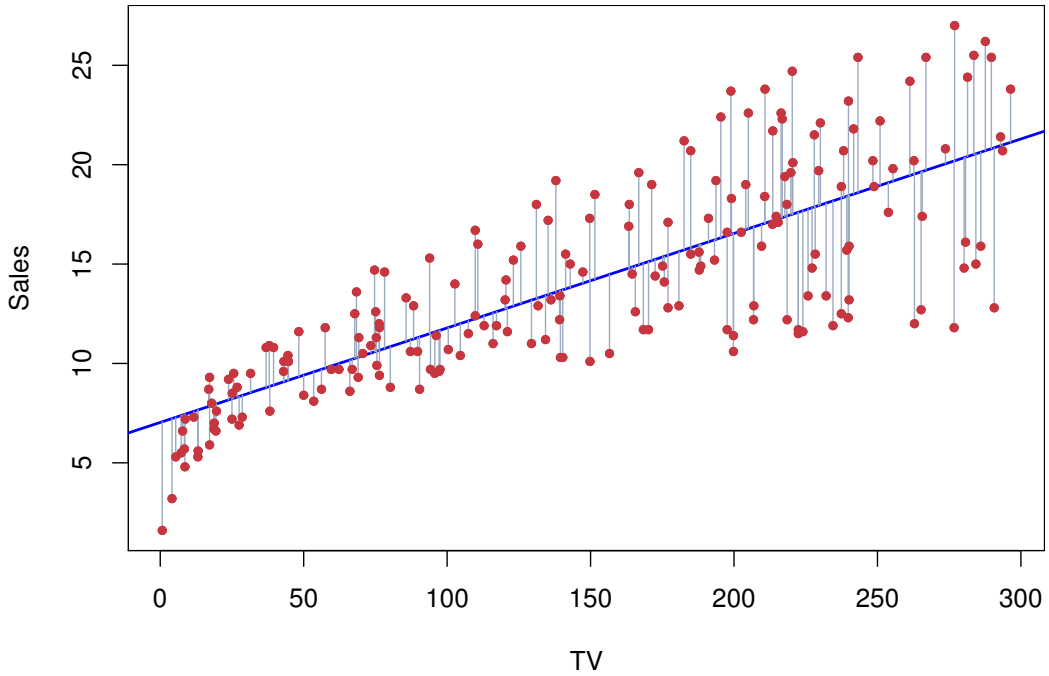
– We may use the simple equalities:

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

to get the final equations

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

How Well Did We Do the Fit?



Now that we have our fit we would like to address few questions about it!

(`We code up the answer to each question in the Example Code 1`)

What is the Confidence Interval for the Coefficients Obtained?

- We can define confidence intervals that the true β_1 and β_0 are in it with 95% confidence
- For $\sigma^2 = \text{var}(\epsilon)$, we define

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

- Hint on derivation: since $\sum_i (x_i - \bar{x})\bar{x} = 0$, we can start with the alternative formulation $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and use the properties stated at the beginning of slides to derive $var(\hat{\beta}_1)$
- The 95% confidence intervals for β_0 and β_1 are

$$\left[\beta_0 - 2\text{SE}(\hat{\beta}_0), \beta_0 + 2\text{SE}(\hat{\beta}_0)\right], \quad \left[\beta_0 - 2\text{SE}(\hat{\beta}_0), \beta_0 + 2\text{SE}(\hat{\beta}_0)\right]$$

`(see the code)`

Is There a Relationship Between x and y?

- We would like to know if there is really a relationship between x and y or if the fit is useless?
- We form a hypothesis testing for β_1 (if it is zero, then x and y are not related):
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$
- Our test statistic is chosen to be $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$
- We look this up in the t-distribution table and find the p-value

(see the code)

If p-value$\leq \alpha$: reject H_0 (accept H_1)
If p-value$> \alpha$: fail to reject H_0

- For the example provided p-value $= 2 \times 10^{-16}$ and we reject H_0 (meaning that x and y are related)

How Well does the Model Explain the Data?

- We can also answer the question of how well our fitted model explains the data by defining another statistic:

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

where TSS is the Total Sum of Squares

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- For general regression problems (not only the simple one) R^2 statistic measures the proportion of variability in y that can be explained by x
- R^2 close to 1 indicates that our model explains a large proportion of the response variability, and R^2 close to zero indicates that our model cannot explain much of the variability in response

`(see the code)`

Multiple Linear Regression

Multiple Linear Regression

– It can be the case that we have multiple features $x_1, \ldots, x_p$ and we would like a fit like

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \epsilon$$

For example: $\text{sales} = \beta_0 + \beta_1.\text{TV} + \beta_2.\text{radio} + \beta_3.\text{newspaper} + \epsilon$

– Suppose that we have n training/response samples $(\boldsymbol{x}^{(1)}, y_1), \ldots, (\boldsymbol{x}^{(n)}, y_n)$ where

$$\boldsymbol{x}^{(1)} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_p^{(1)} \end{pmatrix}, \quad \boldsymbol{x}^{(2)} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_p^{(2)} \end{pmatrix}, \quad \ldots \quad , \boldsymbol{x}^{(n)} = \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_p^{(n)} \end{pmatrix}$$

Multiple Linear Regression

– We would like to minimize the following squared error

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)} \ldots - \beta_p x_p^{(i)} \right)^2$$

– Consider using the following matrices/vectors

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{X}_{n \times (p+1)} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

note that matrix $\boldsymbol{X}$ has the samples along the rows

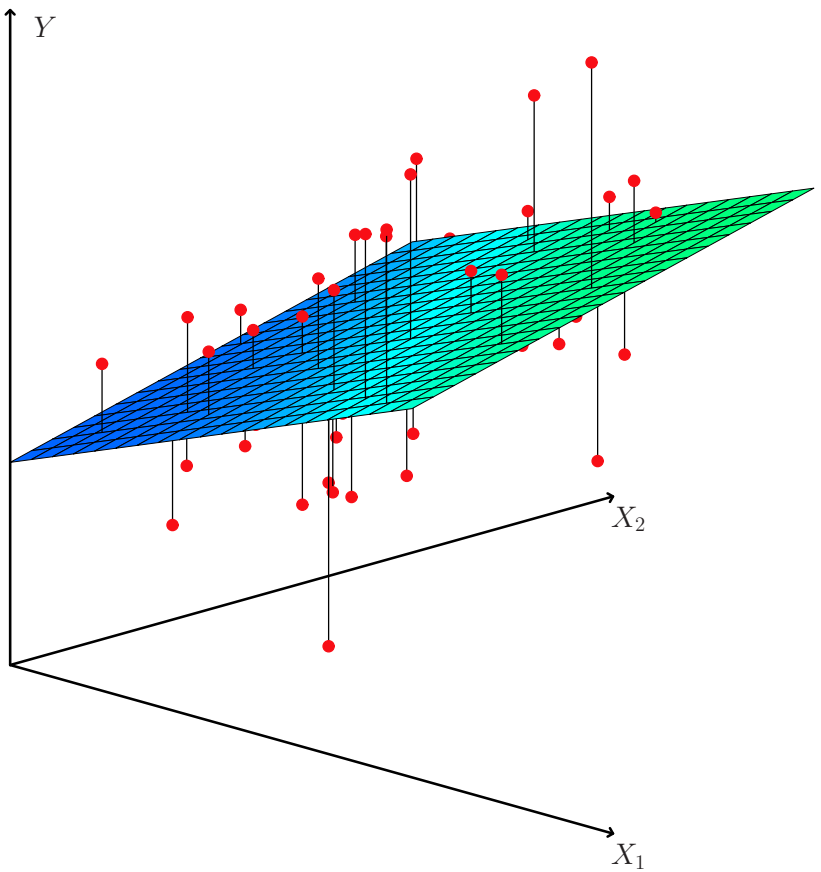
– Then, it is straightforward to see that

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Multiple Linear Regression

– Similar to what we did before we can set $\partial RSS/\partial \boldsymbol{\beta} = 0$ (requires little bit of knowing how to do vector/matrix derivatives) and get

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



Are the Features and Response Related?

- We would like to know if at least one of the features $x_1, \ldots, x_p$ is useful in predicting the response.
- We form a hypothesis testing as follows:
  - $H_0 : \beta_1 = \beta_2 = \ldots = 0$
  - H_1 : at least one β_i is non-zero
- For $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, our test statistic is chosen to be

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

which turns to have an F distribution

If p-value$\leq \alpha$: reject H_0, If p-value$> \alpha$: fail to reject H_0

- We can either use F-distribution tables and find the p-value; or use this rule:*if F is much larger than 1, we reject H_0; if F is very close to 1, we fail to reject H_0*

(see the code)

Now that we have our fit, again we would like to address few questions about it!

(`We code up the answer to each question in the Example Code 2`)

Assessing the P-Values and Correlations Among Features

– Sometimes (most of the times) features are correlated and the contribution of one feature can be taken care of by the others

	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

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

(see the code)

What are the Best Selection of Features?

- As noted sometimes some features can be redundant and we would like to find the best subset of features that predicts well and is not redundant
- In general this problem is "NP-hard" (computationally very hard) and we need to assess 2^p models
- There are some heuristics to do this that we will see later:
  - Forward selection: model p regressions each with only one feature, pick the one with least RSS, repeat it with selected feature and combination of others, ...
  - Backward selection: Start with all features and remove variable with largest p-value, run a new regression, remove variable of largest p-value, ...

How Can We Handle Categorical Features?

- Sometimes our features do not take numerical values, instead they take categorical values
- **Example**: In a regression problem we have a feature called `ethnicity`, which takes possible values of `Asian, Caucasian, African-American`
- We can introduce 2 dummy variables (features) e_A, e_C
  - $e_A = 1,\ e_C = 0$ if Asian
  - $e_A = 0,\ e_C = 1$ if Caucasian
  - $e_A = 0,\ e_C = 0$ if African-American
- Basically, for every categorical feature that has L levels, we need to define $L - 1$ dummy variables

Can We Only Fit Flat Curves with Linear Regression?

– Based on the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \epsilon$$

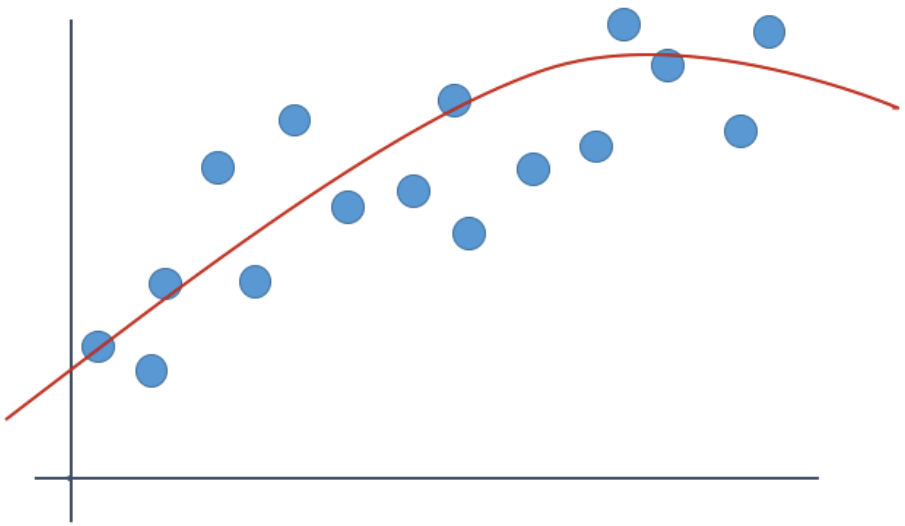
one might get the impression that linear regression is only good for fitting flat surfaces (linear manifolds)

– If we include powers of a feature, e.g., $x_1, x_1^2, \ldots$ or cross terms between the features, e.g., $x_1 x_2, x_2 x_3 x_5$, etc, then we can also fit nonlinear surfaces

– Of course knowing what powers or what cross terms to include in the feature list is not always clear

Can We Only Fit Flat Curves with Linear Regression?

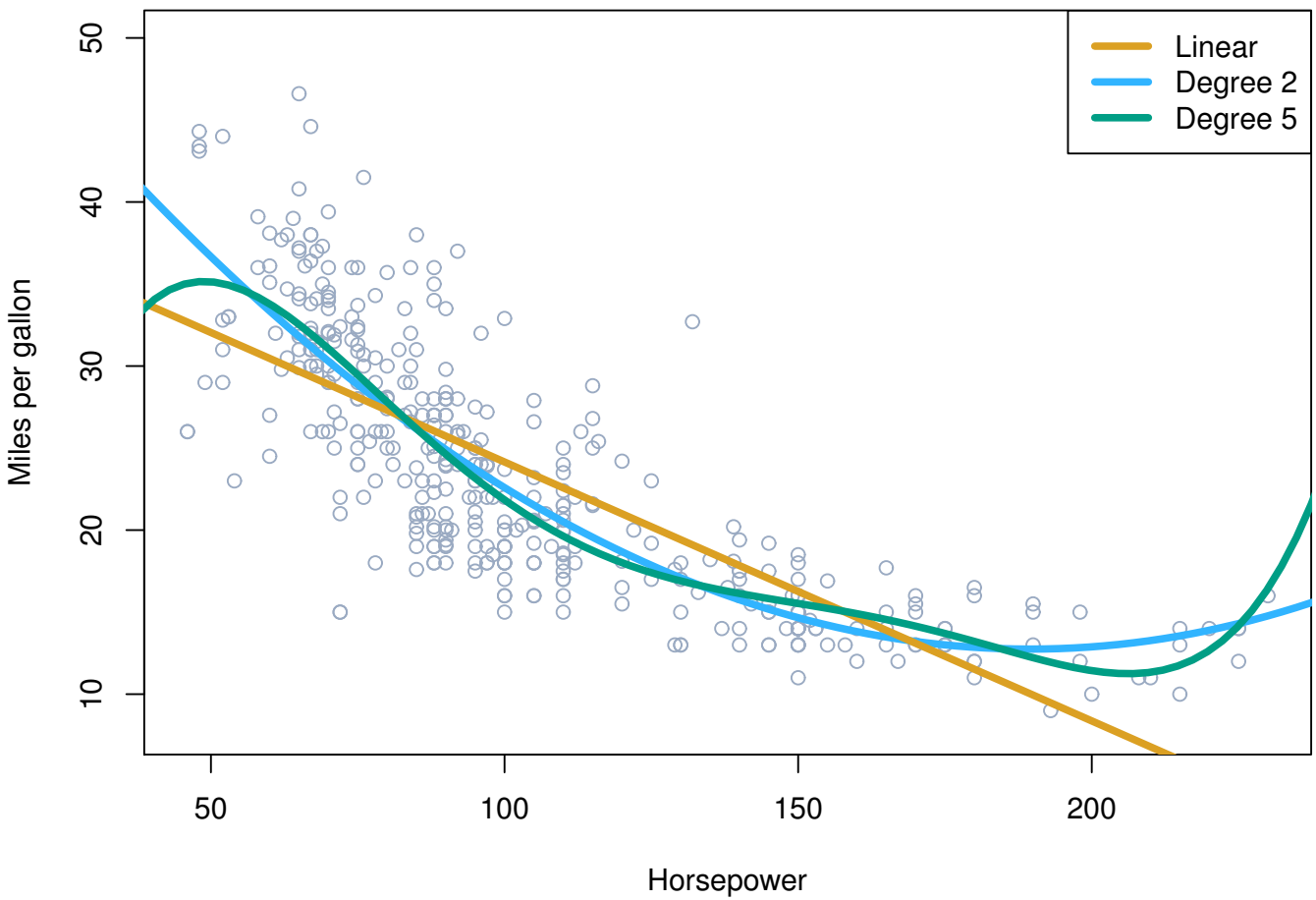
– **Example**: For a problem wtih only one feature, we have a set of point that look lying on a parabola, we use the regression $y = \beta_0 + \beta_1 x + \beta_2 x^2$



Can We Only Fit Flat Curves with Linear Regression?

– **Example**: Regressing Mile per Gallon in terms of the Horse Power

$$\texttt{mpg} = \beta_0 + \sum_{i=1}^{p} \beta_i (\texttt{horsepower})^i$$



Questions?

References

https://www.alsharif.info/iom530, 2013.

J. Friedman, T. Hastie, and R. Tibshirani.
The elements of statistical learning.
Springer series in statistics, 2nd edition, 2009.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/linear_regression.pdf, 2013.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
An introduction to statistical learning: with applications in R, volume 112.
Springer, 2013.