

# Applied Regressions Midterm Review

5 weeks in an hour

Connor Dowd

Liberal borrowing slides from Panos Toulis

University of Chicago's Booth School of Business

October 29, 2018

## 1 Basics

Covariance, Correlation, Lines, Residuals, Least Squares

## 2 Inference

Significance, Model, Sampling Distribution, CIs, P-values

## 3 MLR

Interpretation, Categoricals, Interactions, Polynomials

## 4 Diagnostics

Assumptions, Leverage, Constant Variance, Non-linearity, Multi-collinearity

## 5 Model Selection

F-test, Bonferonni, AIC, Stepwise, LASSO, CV

# Week 1: Regression Basics

- ① What is Regression?
- ② Covariance, Correlation
- ③ Conditional Distribution
- ④ Lines: Slopes and Intercepts
- ⑤ Residuals
- ⑥ Least Squares

## Regression: What is it?

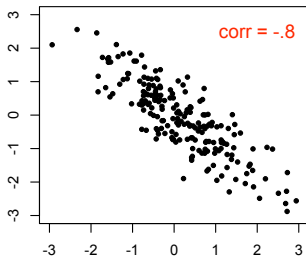
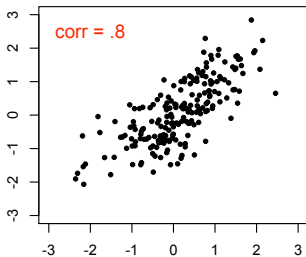
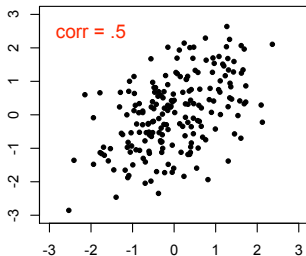
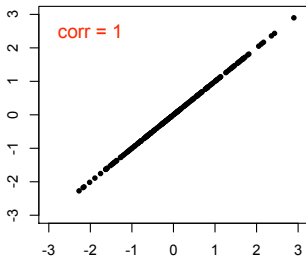
- Investigate relationship between  $Y$  (response) and  $X_1, X_2, \dots$  (explanatory variables) is expressed as:

$$Y = f(X_1, X_2, \dots) + \varepsilon.$$

- $\varepsilon$  is noise with zero mean: if we fix the values of  $X_1, X_2, \dots$ , then

$$E(Y|X_1 = x_1, X_2 = x_2, \dots) = f(x_1, x_2, \dots).$$

Correlation  $\rho_{xy}$  :  $\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)} \in [-1, +1]$ .



Regression improves on correlation by allowing us to make predictions.

## Option 2: conditional distributions

The conditional distribution of  $Y$  given  $X$  contains much more information than correlation. And can be used for prediction!

**Example:** Suppose we see (old) house price data comprised of pairs (price  $\times 10^3$  dollars, size  $\times 10^3$  sq.ft.) as such:  
(100, 1.5), (85, 1.4), (75, 1.0), (90, 1.5).

*What is your prediction for a 1,500 sq.ft. house?*

## Option 2: conditional distributions

The conditional distribution of  $Y$  given  $X$  contains much more information than correlation. And can be used for **prediction**!

**Example:** Suppose we see (old) house price data comprised of pairs (price  $\times 10^3$  dollars, size  $\times 10^3$  sq.ft.) as such:  
(100, 1.5), (85, 1.4), (75, 1.0), (90, 1.5).

*What is your prediction for a 1,500 sq.ft. house?*

- Good guess:  $(100 + 90)/2 = \$95,000$ .
- In our guess, we used (a summary of) the conditional distribution of  $Y$  given  $X$ .

## Why is the conditional distribution not enough?

**Example:** Suppose we see house price data comprised of pairs (price  $\times 10^3$  dollars, size  $\times 10^3$  sq.ft.) as such:  
(100, 1.5), (85, 1.4), (75, 1.0), (90, 1.5).

*What is your prediction for a **1,300 sq.ft** house?*



## Why is the conditional distribution not enough?

**Example:** Suppose we see house price data comprised of pairs (price  $\times 10^3$  dollars, size  $\times 10^3$  sq.ft.) as such:  
(100, 1.5), (85, 1.4), (75, 1.0), (90, 1.5).

*What is your prediction for a **1,300 sq.ft** house?*

- Good guess?
- No data for a 1,300 sq.ft. house.
- Our prediction **has to** rely on a key intuition:  
*“A 1,300 sq.ft house cannot be that different from a 1,400 sq.ft or a 1,200 sq.ft. house”.*
- We need a **model** that will allow **extrapolation**!

## Linear Regression Model

In a linear regression, that

extrapolation will be linear in nature.

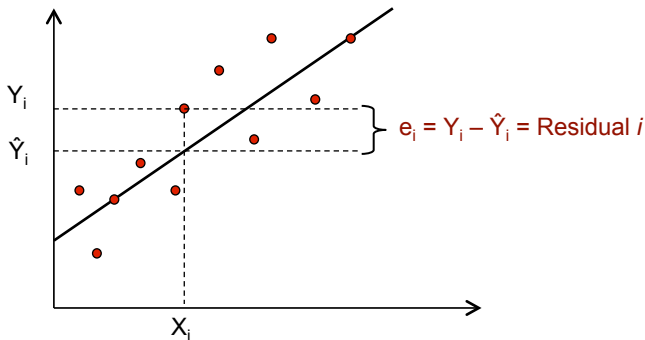
Recall that the equation of a line is:

$$\hat{Y} = b_0 + b_1X,$$

where  $b_0$  is the **intercept** and  $b_1$  is the **slope**.

In the house price example

- our “eyeball” line had  $b_0 = 35$ ,  $b_1 = 40$ .
- **predict** the price of a house when we know only size
  - $35 + 40 \times 1.3 = 87,000$  dollars.
- The intercept value is in units of  $Y$  (\$1,000).
- Slope is in units of  $Y$  *per* units of  $X$  (\$1,000/1,000 sq ft).



The line is our predictions or **fitted values**:  $\hat{Y}_i = b_0 + b_1 X_i$ .

The **residual**  $e_i$  is the discrepancy between the **fitted**  $\hat{Y}_i$  and **observed**  $Y_i$  values, i.e.,

$$e_i = Y_i - \hat{Y}_i.$$

- Note that we can write  $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$ .

## Least squares

A reasonable goal is to minimize the size of *all* residuals:

- Trade-off between moving closer to some points and at the same time moving away from other points.

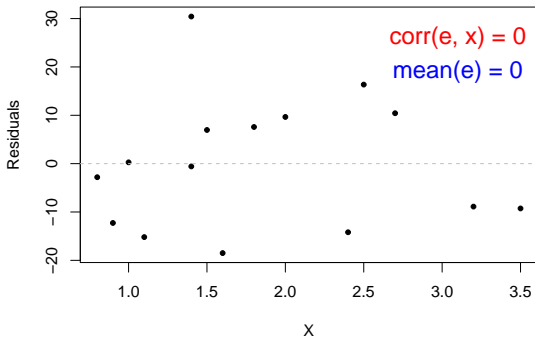
Since some residuals are positive and some are negative, we need one more ingredient.

- $|e_i|$  treats positives and negatives equally.
- So does  $e_i^2$ , which is easier to work with mathematically.

Least squares chooses  $b_0$  and  $b_1$  to minimize  $\sum_{i=1}^n e_i^2$ .

The residuals from least squares regression are “stripped of all linearity”.

```
> plot(size, reg$fitted-price, pch=20, xlab="X", ylab="Residuals")
> text(x=3.1, y=26, col=2, cex=1.5,
+      paste("corr(e, x) =", round(cor(size, reg$fitted-price),2)))
> text(x=3.1, y=19, col=4, cex=1.5,
+      paste("mean(e) =", round(mean(reg$fitted-price),0)))
> abline(h=0, col=8, lty=2)
```



# Week 2: Inference

- ① Sampling Distribution
- ② Statistical Significance
- ③ Regression Model
- ④ Inferential Setup
- ⑤ Confidence Intervals
- ⑥ P Values
- ⑦ Bootstrap?
- ⑧ Prediction vs Confidence Intervals

# Sampling Distributions

All possible datasets  $(X, Y)$  given  $X$  and  $n$  data points.



These different possible distributions create the sampling distributions. But we don't have those datasets! How to proceed?

- 1 Assume a **model** that generates them! Use  $b_0, b_1$  to make **inference** about model.
  - (coming next) Simple Linear Regression model (SLR).
- 2 Pull yourself up by the **bootstrap**.

# Simple linear regression (SLR)

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- It is a **model**, so we are *assuming* this relationship holds for some **fixed but unknown** values of **parameters**:

$$\beta_0, \beta_1, \sigma^2.$$

- The error  $\varepsilon$  is **independent**, **additive**, idiosyncratic noise. Implies regression model:

$$E(Y|X) = \beta_0 + \beta_1 X.$$

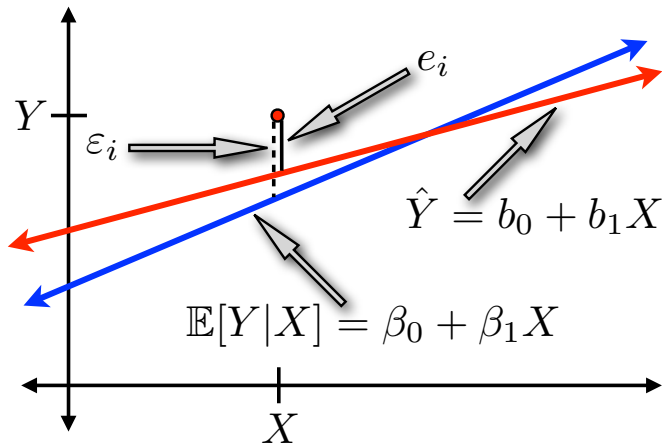
▷ *SLR* (“*simple*” = *only one X*; “*linear*” = *linear in X*).



**IMPORTANT!**  $\beta_0$  is not  $b_0$ ,  $\beta_1$  is not  $b_1$ , and  $\varepsilon$  is not  $e$ .

True regression line is **fixed** (but unknown).

Least-squares line **changes** wrt observed data(=random).



(We use Greek letters to remind us.)



## Inference setup

model	least squares	comments
$Y = \beta_0 + \beta_1 X + \varepsilon$	$\hat{Y} = b_0 + b_1 X$	
$\beta_0$	$b_0$	intercept
$\beta_1$	$b_1$	slope
$\varepsilon$	$e = Y - \hat{Y}$	noise/residuals (obs-pred)
$\sigma^2$	$s^2$ or $\hat{\sigma}^2$	noise error

- **Goal:** Do inference on  $\beta_0, \beta_1, \sigma^2$ .
- Intuitively: use  $b_0, b_1, \sum_i e_i^2$ , respectively.
- The key concept for inference is **sampling distribution**.

# Sampling distribution of $b_1$ (theory)

It turns out that  $b_1$  is normally distributed:

$$b_1 \sim \mathcal{N}(\beta_1, \sigma_1^2).$$

- $b_1$  is unbiased:  $\mathbb{E}[b_1] = \beta_1$ , mean centered at true value.  
*What is that expectation over?*
- The sampling variance  $\sigma_1^2$  determines estimate precision:

$$\sigma_1^2 = \text{Var}(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1)s_x^2}.$$

It depends on three factors:

- 1 sample size ( $n$ ) – *sample size increases,  $b_1 \rightarrow \beta_1$ ;*
- 2 error variance ( $\sigma^2$ ) – *more noise means less information;*
- 3  $X$ -spread ( $s_x$ ) – *more var in  $X$  means more information!*

# Sampling distribution of $b_0$

The intercept is also **normal** and **unbiased**:

$$b_0 \sim \mathcal{N}(\beta_0, \sigma_0^2).$$

The variance is given by:

$$\sigma_0^2 = \mathbb{V}ar(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2} \right).$$

– *usually more variance in intercept than in slope.*

- The sampling variance depends on **four factors**:
  - 1 sample size ( $n$ ),
  - 2 error variance ( $\sigma^2$ ) and  $X$ -spread ( $s_x$ ), and
  - 3 squared mean of explanatory variables,  $\bar{X}^2$ .

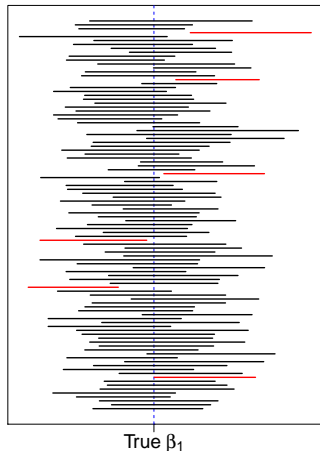
# Confidence intervals : illustration

The “frequentist” interpretation.

$$\mathbb{P}\left[\beta_1 \in (b_1 \pm 2\sigma_1)\right] = 95\%$$

**What is probability over?**

- CI determines a range of **plausible values** for unknown parameter.
- Center of range = **point estimate**.
- Length of range = **uncertainty**.



# p-values tell us more

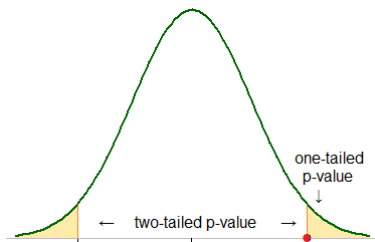
Test  $\beta_1 = 0$ . Figure is standardized distribution under null.

$$T_1 = \frac{b_1 - 0}{s_1} \sim_{H_0} t_{n-2}.$$

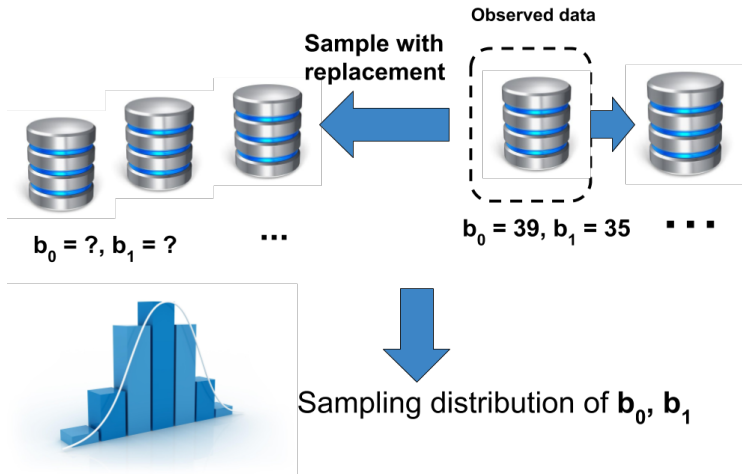
- $p$ -value = area of shaded region. <sup>s1</sup>Using R:

```
T1 = (b1 - 0) / s1      # test statistic  
pval = 2 * pt(-abs(T1), df=n-2)
```

- This is a two-sided  $p$ -value, which are usually used with symmetric distributions.
- Lower  $p$ -value = stronger evidence to reject  $H_0$ .
- For level 0.05: Reject if  $p\text{-value} \leq .05$ .



# Bootstrap: illustration

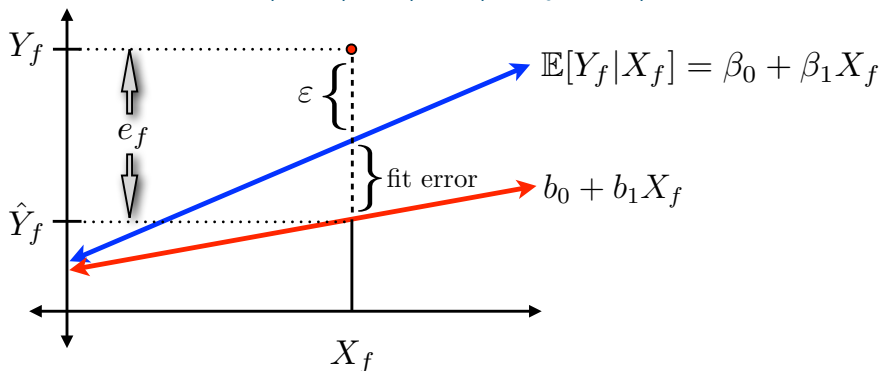




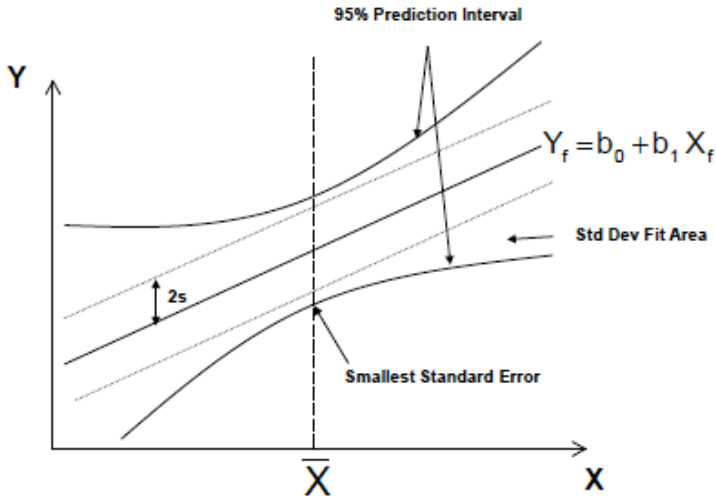
**Prediction Intervals** Suppose  $\hat{Y}_f$  is our prediction for  $Y$  based on  $X_f$ .

If we use  $\hat{Y}_f$ , our **prediction error** has **two** pieces

$$e_f = Y_f - \hat{Y}_f = Y_f - b_0 - b_1 X_f.$$



⇒ The prediction (conf.) interval needs to **widen away from  $\bar{X}$**



# Week 3: Multiple Regression

- ① Interpretation
- ② Categoricals: Dummies & Factors
- ③ Interactions
- ④ Polynomials

# The MLR Model

The MLR model is linear but with **more** variables:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

or equivalently

$$Y|X_1, \dots, X_d \stackrel{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d, \sigma^2).$$

Recall the key assumptions of our linear regression model:

- (i) The conditional mean of  $Y$  is **linear** in the  $X_j$  variables.
- (ii) The additive errors (deviations from line) are **iid**:
  - **i**ndependent from each other,
  - **i**dentically distributed; Normal, in particular.

## Interpretation

Interpretation of model parameters can be extended from SLR:

$$\beta_j = \frac{\partial \mathbb{E}[Y|X_1, \dots, X_d]}{\partial X_j}.$$

- $\partial$  is from calculus and means “change in”
- Holding all other variables constant,  $\beta_j$  is the change in the regression function (=conditional expectation of  $Y$ ) per unit change in  $X_j$ .

# Reading Model Summary Information

Call:

```
lm(formula = y ~ x + z)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.4433	-0.6795	-0.0343	0.6780	3.4987

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.002768	0.032108	-0.086	0.931
x	2.999396	0.032907	91.147	<2e-16 ***
z	0.503753	0.031702	15.890	<2e-16 ***

---

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

Residual standard error: 1.015 on 997 degrees of freedom

Multiple R-squared: 0.8962, Adjusted R-squared: 0.896

F-statistic: 4304 on 2 and 997 DF, p-value: < 2.2e-16

# MLR topic 1: Categorical variables

Allows to **drill down** to subpopulations.

In R a categorical variable is called **factor**.

The simplest factor is **binary** (also known as **dummy variable** or **indicator**); e.g.,

- temporal effects (1 if Holiday season, 0 if not);
- spatial (1 if in Midwest, 0 if not).

We use special notation:

- $\mathbb{1}_{[X=r]} = 1$  if  $X = r$ , 0 if  $X \neq r$ .
- Use  $R - 1$  dummies for a factor  $X$  with  $R$  possible levels.

## Example: Credit Data

$$\mathbb{E}[\text{Balance}|\text{Student}] = \beta_0 + \beta_1 \mathbb{1}_{[\text{Student}=\text{yes}]}.$$

Model says:

- $\mathbb{E}[\text{Balance}|\text{Student} = \text{no}] = \beta_0.$
- $\mathbb{E}[\text{Balance}|\text{Student} = \text{yes}] = \beta_0 + \beta_1.$

Easy in R:

```
> summary(lm(Balance ~ Student))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	480.37	23.43	20.50	< 2e-16 ***
StudentYes	396.46	74.10	5.35	1.49e-07 ***

---

\* Expected balance for non-students = \$480.37.

\* Expected balance for students = \$480.37 + 396.46 = 876.83.



## MLR topic 2: Interactions

Suppose  $X_1$  is a dummy variable.

$$Y = \beta_0 + \beta_1 \mathbb{1}_{\{X_1=1\}} + \beta_2 X_2 + \beta_3 \mathbb{1}_{\{X_1=1\}} X_2 + \cdots + \varepsilon.$$

$$\frac{\partial \mathbb{E}[Y|X_1 = 0, X_2, \dots]}{\partial X_2} = \beta_2. \quad \frac{\partial \mathbb{E}[Y|X_1 = 1, X_2, \dots]}{\partial X_2} = \beta_2 + \beta_3.$$

- 
- \* The model includes **interaction** of  $X_1$  with  $X_2$  ( $\beta_3$  term).
  - \* Slope of  $X_2$  effectively depends on  $X_1$ .

## Interactions with R are easy!

```
> reg = lm(Balance ~ Rating * Student)
> summary(reg)
...
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-423.37117	26.14584	-16.193	< 2e-16	***
Rating	2.54543	0.06747	37.728	< 2e-16	***
StudentYes	311.99938	85.86752	3.633	0.000316	***
Rating:StudentYes	0.24609	0.22372	1.100	0.272002	

- For **non-students**, the SLR and LS models are, resp.:

$$\text{Balance} = -423.37 + 2.54 \times \text{Rating} + \varepsilon.$$

$$\mathbb{E}[\text{Balance}|\text{Rating}] = -423.37 + 2.54 \times \text{Rating}.$$

- For **students**, the SLR and LS models are, resp.:

$$\text{Balance} = -111.37 + 2.79 \times \text{Rating} + \varepsilon.$$

$$\mathbb{E}[\text{Balance}|\text{Rating}] = -111.37 + 2.79 \times \text{Rating}.$$

## MLR topic 4: Polynomial regression

Polynomial regression is special case of variable transformation.

We simply take powers of single variable  $X$ :

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m.$$

You can fit any mean function if  $m$  is big enough.

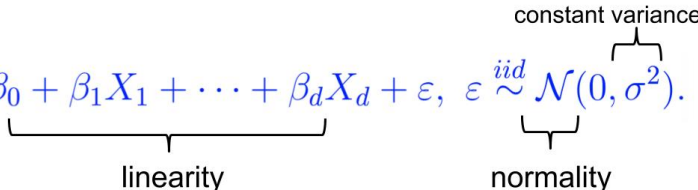
- Usually,  $m = 2$  does the trick.

# Week 4: Diagnostics

- 1 Model Assumptions
- 2 Leverage (outliers?)
- 3 Q-Q plots
- 4 Constant Variance
- 5 Non-linearity
- 6 Log-log (and elasticities)
- 7 Multi-collinearity

# Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

  
linearity                      normality                      constant variance

We are going to **check** the following model assumptions:

- ① Errors are identically distributed as **normals**...
  - Topic A: **outliers**.
- ② ...with **constant** variance  $\sigma^2$ .
- ③  $\mathbb{E}(Y|X_1, \dots, X_d)$  is **linear** in  $X_1, \dots, X_d$ .
  - Topic B: **log-log model**.
  - Topic C: **multicollinearity**.

## Why bother?

If the model assumptions do not hold then

- prediction can be **biased**;
- standard errors and confidence intervals may be **wrong**;
- could lead to model **improvements** (e.g., transformations).

Plots of residuals  $e = Y - \hat{Y}$  are our **#1 tool**.

## Theoretical sampling distribution of $e_i$

If MLR model is true:

$$e_i \sim \mathcal{N}(0, \sigma^2[1 - h_i]), \quad h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)s_X^2}.$$

The  $h_i$  term is referred to as the  $i^{th}$  observation's **leverage**:

- It is that point's share of the data ( $1/n$ ) plus its proportional contribution to variability in  $X$ .
- Leverage is **minimum** when  $X_i = \bar{X}$ .

---

\* Notice that if  $n =$  very large, the residuals  $e_i$  “obtain” the same distribution as the errors  $\varepsilon_i$ , i.e., approximately  $e_i \sim \mathcal{N}(0, \sigma^2)$ .

## Standardized residuals

We can use the familiar standardization trick:

$$r_i = \frac{e_i}{\sigma \sqrt{1 - h_i}} \sim \mathcal{N}(0, 1).$$

These transformed  $e_i$ 's are called the **standardized** residuals.

- Replace  $\sigma^2$  with  $s^2$  (**studentized** residuals).
- In R use `rstudent(fit)` for studentized residuals.

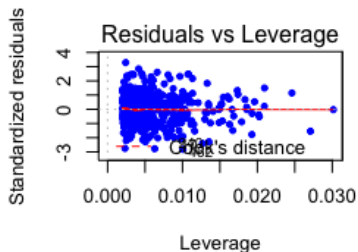
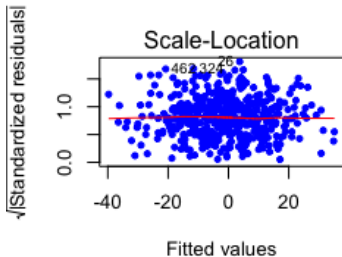
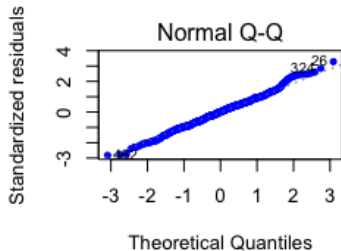
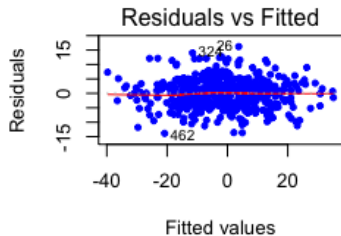
---

▷ Great result. Now we can check deviations from MLR assumptions by looking at whether studentized residuals look normal (or t-distributed).




# R suite of diagnostics

>



## What to do about non-constant variance?

No easy solutions. There are generally three things to do:

- Transformations! (coming later)
- Subpopulation analysis.
- Model  $e^2 \sim X$
- **Robust standard errors** (“sandwich” estimator ).

```
> library(sandwich)
> # classical std errors:
> summary(fit)$coefficients[, 2]
(Intercept)          x1          x2
  0.1399635   0.1413205   0.1404328
> # robust std errors
> sqrt(diag(vcovHC(fit)))
(Intercept)          x1          x2
  0.1406091   0.1964196   0.1900342
```

- 
- \* Before using: `install.packages("sandwich")`.
  - \* Robust error takes into account heteroskedasticity.

### 3. Non-linearity check

- Probably the most common model violation.
- Linearity may hold under *some* transformation of  $X$  or  $Y$ .
- Example is polynomial regression (e.g.,  $Y = X + X^2 + \dots$ ).
- In general, think about in what scale you expect linearity.

- 
- \* Diagnostics so far may reflect failure of linearity.
  - \* No general test for that.
  - \* More interesting to focus on general classes of models.

## Topic B: the log-log model

- If we expect  $Y \approx X^\beta$  then use the log-log model:

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon.$$

- $\log(Y)$  is a very common transformation;
  - e.g., if  $Y$  has only positive values (e.g. sales) or is a count (e.g. # of customers).
- $\log(X)$  is also common to reduce spread in  $X$ .

Recall that

- $\log$  is always natural log, with base  $e = 2.718\dots$ , and
- $\log(ab) = \log(a) + \log(b)$ ,
- $\log(a^b) = b \log(a)$ .

# Price elasticity

In marketing, the slope coefficient  $\beta_1$  in the regression

$$\log(\text{sales}) = \beta_0 + \beta_1 \log(\text{price}) + \varepsilon,$$

is called **price elasticity**.

The model is **multiplicative**:  $\mathbb{E}[\text{sales}|\text{price}] = A * \text{price}^{\beta_1}$ .

## Interpretation:

$\beta_1 =$  % change in **sales** per 1% change in **price**.

## Topic C. Multicollinearity

**Multicollinearity** refers to strong linear dependence between some of the variables in MLR model.

The usual marginal effect interpretation is lost:

- change in one  $X$  variable leads to **change in others**.

Coefficient standard errors will be **large**, such that multicollinearity leads to large uncertainty about the  $b_j$ 's.

Multicollinearity is not a big problem in and of itself, you just need to know that it is there.

If you recognize multicollinearity:

- Understand that the  $\beta_j$  are **not true** marginal effects.
- Consider dropping variables to get a more simple model.
- Expect to see big standard errors on your coefficients (i.e., your coefficient estimates are **unstable**).
  - Can check with `vif(fit)`. Problem if numbers  $> 5$ .
- Use machine learning tools...
- Remember that significance is harder to assess. Recall that in the Advertising data newspaper was significant on its own, but not in presence of radio.

# Week 5: Model Selection

- 1  $R^2$ ?
- 2 F-Test
- 3 Bonferonni
- 4 Model Building?
- 5 AIC
- 6 Stepwise w/ AIC
- 7 LASSO
- 8 Cross-Validation
- 9 Bias-Variance



## Weakness of $R^2$

$$R^2 = \text{cor}^2(\hat{y}, y)$$

- $R^2$  only depends on **model fit** but not **model size**.
- Always improves if we add more variables...
- ...even if those variables **are just noise**.

```
> ad$bogus = rnorm(length(ad$TV))  
> reg = lm(sales ~ TV + bogus, data=ad)  
> summary(reg)$r.sq  
[1] 0.6181036 # R^2 improves but 'bogus' is just noise
```

---

\* We need a more careful way of assessing  $R^2$  increase!

## Akaike information criterion (AIC)

Assesses the quality of a model (like  $R^2$ , but better!):

$$\text{AIC} = 2 * \text{model\_size} + 2 * \text{model\_error}.$$

- Smaller AIC is better.
- Extends beyond MLR. Has better theory than  $R^2$ .

# Basic Model Selection with F-test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{d-1} = 0.$$

$$H_1 : \text{at least one } \beta_j \neq 0.$$

Use the following test statistic ( $F$ -value):

$$f = \frac{R^2/(d-1)}{(1-R^2)/(n-d)}.$$

- This  $F$ -test tries to formalize the idea of a **big**  $R^2$ .
- Someone figured out the distribution of  $f$  if  $H_0$  is true.
- So, if observed  $f$  is big wrt to that distribution then regression is “worthwhile”:

The test is contained in the R [summary](#) for any MLR fit.

## Extended Model Selection with F-test

Suppose that we have  $k$  variables in the base model and we are thinking off adding  $d - k$  new variables.

Essentially, we are asking: “Add more variables to the model?”

$$H_0 : Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1} + \varepsilon.$$

$$H_1 : Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1} + \beta_k X_k + \dots + \beta_{d-1} X_{d-1} + \varepsilon.$$

An equivalent way to set this hypothesis:

$$H_0 : \beta_k = \beta_{k+1} = \dots = \beta_{d-1} = 0.$$

$$H_1 : \text{at least one } \beta_j \neq 0 \text{ for } j \geq k.$$

Use `anova(base.model, large.model)` and look at p-value.

## Bonferroni correction

- F-test tests all betas together, if you want to test jointly, you need to use Bonferonni corrections.
- If you test  $m$  nulls then reject when  $p\text{-value} < 0.05/m$ .

## Universe of variables

The universe of variables is **LARGE!**

- includes all possible variables that you think might have a linear effect on the response,
- ...and all squared terms ...and all interactions ....

**You** decide on this universe through your experience and domain knowledge (and are limited by data availability).

# 1. Forward stepwise regression

- 1 Run  $Y \sim X_j$  for each variable, then choose the one leading to the smallest AIC to include in your model.
- 2 Given you chose covariate  $X^*$ , now run  $Y \sim X^* + X_j$  for each remaining  $j$ , again based on smallest AIC.
- 3 Repeat this process until none of the expanded models lead to smaller AIC than the previous model.

This is called “forward stepwise regression”.

## Notes on stepwise regression

- You **can't be** absolutely sure you've found the best model.
- Forward stepwise regression follows **greedy approach** and is going to miss groups of variables that are only influential together.
- Usually leads to **dense models**, which is not ideal.

It's not perfect, but it is pretty handy.



# LASSO

We're going to skip most details here. The short version is:

$$\min \left\{ \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^p |b_j| \right\}$$

This expression has three pieces:

- ① First term is standard RSS.
- ②  $\sum_{j=1}^p |b_j|$  measures the model's complexity
  - LASSO fit get you lots of  $b_j = 0$ .
  - Final model is variables with  $b_j \neq 0$ .
- ③  $\lambda$  determines how important the penalty is;
  - Choose by cross-validation (R does all the work).
- This leads to sparse solutions and scales very well.

### 3. Cross-validation

How does LASSO choose the penalty? ( $\lambda$  value)

- Cross-validation (CV).

CV is a model assessment tool like  $R^2$  and AIC. But it is also very different:

- It directly estimates the model's **generalization error**.
- Does not rely on model assumptions/approximations.
- Very **easy** to implement (could be **slow** though).

---

\* Next: **in-sample** and **out-of-sample error** (generalization).

# Cross-validation

In  $k$ -fold cross-validation:

- Split the data in  $k$  folds of equal size.
- For every fold  $i = 1, 2, \dots, k$ :
  - Train in all folds other than  $i$ .
  - Calculate  $\text{MSE}_i$  at fold  $i$ .
- Cross-validation estimate of generalization error:

$$\text{CV} = \frac{1}{k} \sum_{i=1}^k \text{MSE}_i.$$

Very powerful way to **detect** and **prevent** overfitting!

# Summary

	assessment	building	selection
Classical	$R^2$ -inc. with model size	Nested models -cannot scale	F-test -MLR assumptions
Modern	AIC +Generally useful	forward regression +Automated,-Greedy	AIC -May be inaccurate
Hot	CV error +Easy to implement	Regularization +Sparse models	CV +Broadly applicable

# Good luck!

- Questions?