

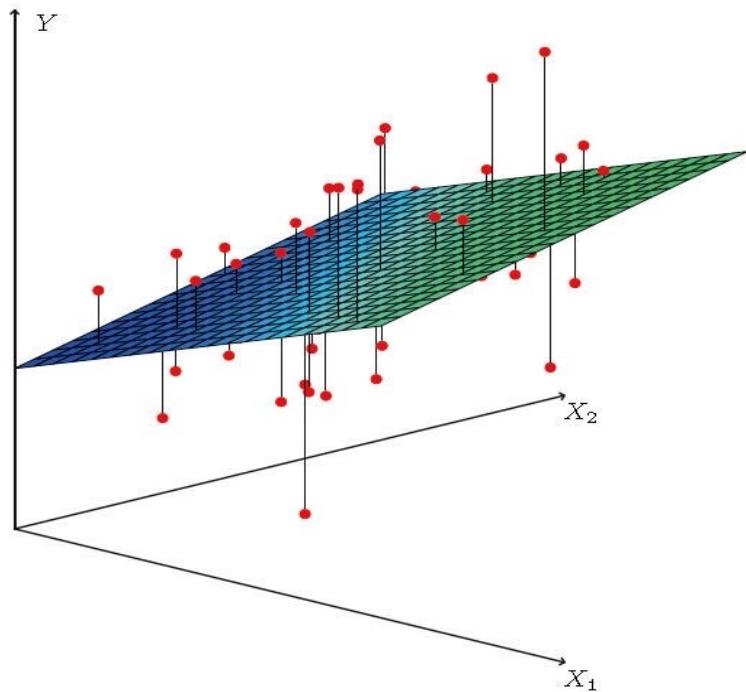
EE 559: Machine Learning I

University of Southern California

M. R. Rajati, PhD

Lesson 2

Linear Regression

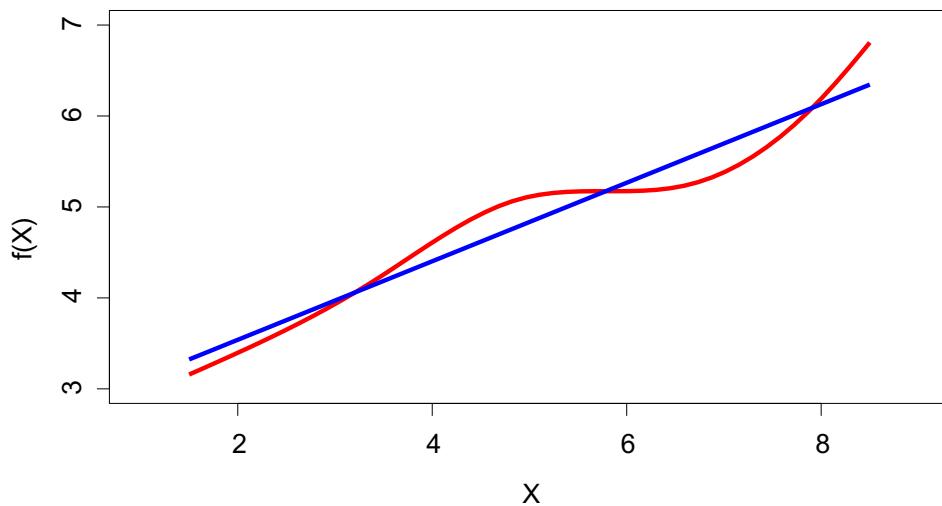


Linear Regression

- 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.

Linear regression

- 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 shown on the next slide. Questions we might ask:

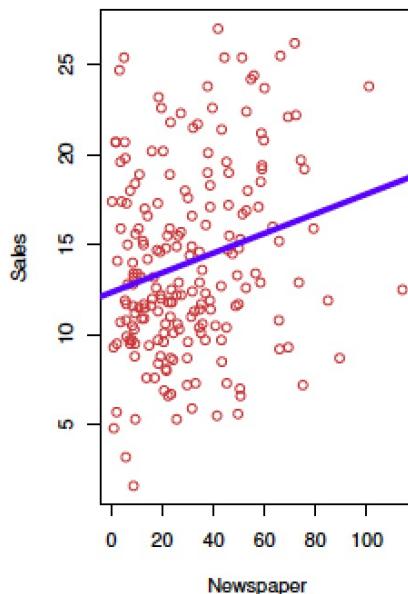
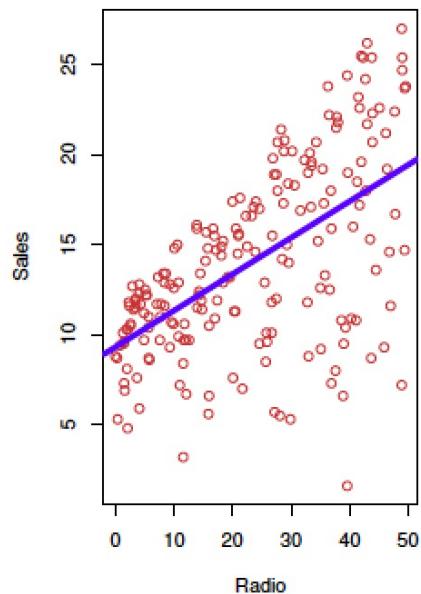
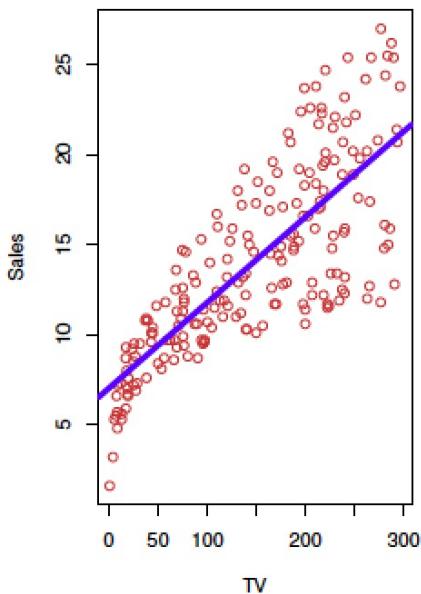
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?

Linear regression for the advertising data

Consider the advertising data shown on the next slide. Questions we might ask:

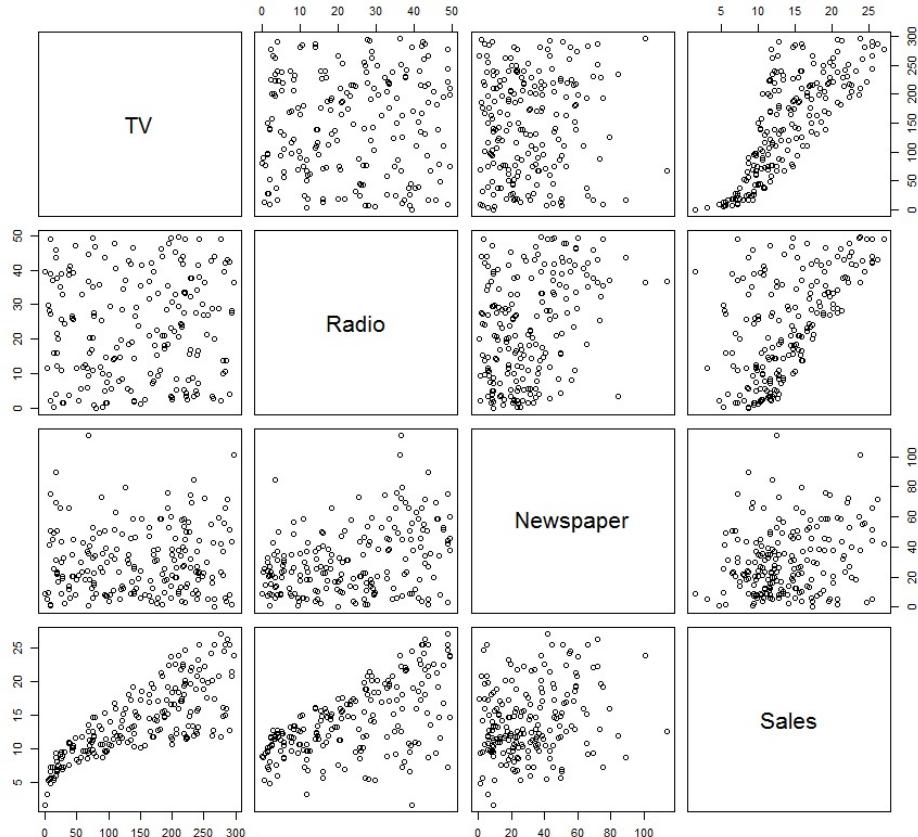
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Advertising data



Case 1: Advertisement Data

```
Advertising=read.csv("http://www-
bcf.usc.edu/~gareth/ISL/Advertising.csv", header=TRUE);
newdata=Advertising[,-1]
fix(newdata)
View(newdata)
names(newdata)
pairs(newdata)
```



Simple linear regression using a single predictor X .

- We assume a model

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

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ε is the error term.

Simple linear regression using a single predictor X .

- 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*

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 $\text{RSS} = e_1^2 + e_2^2 + \dots + e_n^2$ or equivalently as

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

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

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

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

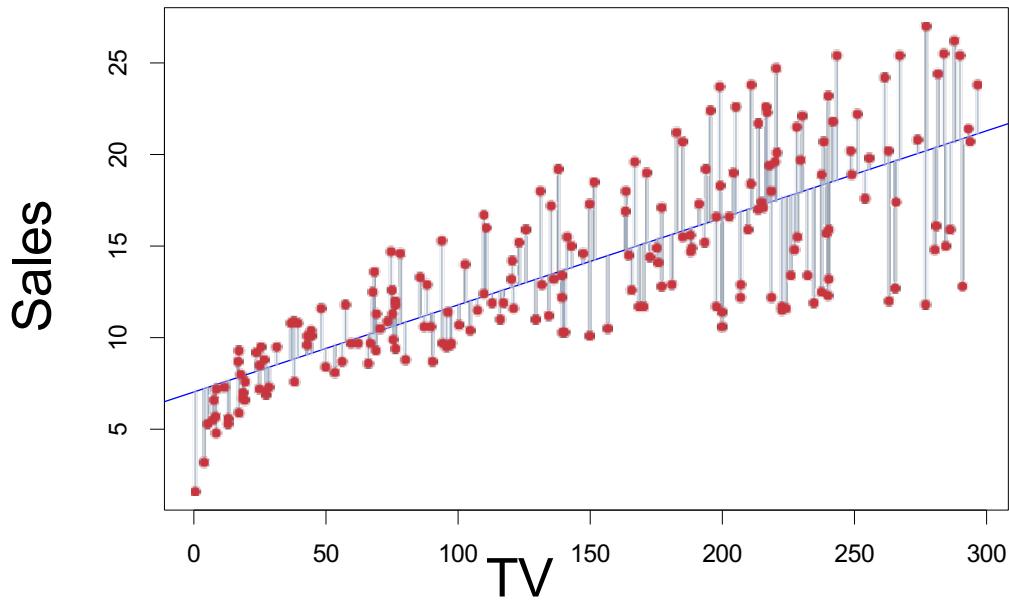
where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

Estimation of the parameters by least squares

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

Estimation of the parameters by least squares

Example: advertising data



The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Assumptions

$$Y \approx f(X) + \varepsilon$$

$$\cdot y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$f(X) = \beta_0 + \beta_1 X$$

where ε_i 's are independent and

$$E[\varepsilon_i] = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

identically
distributed

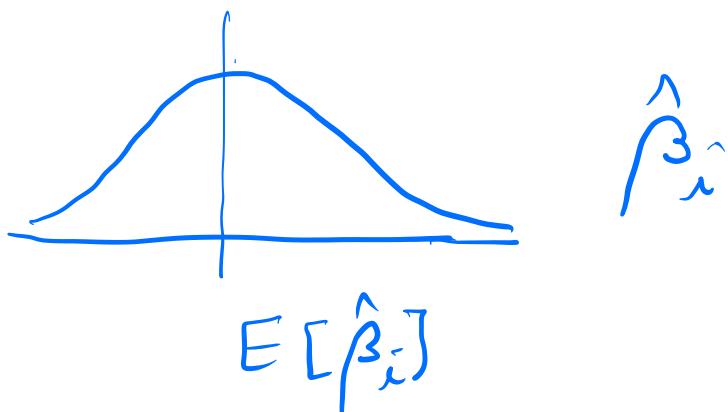
ε_i and x_i are independent

Normality is not needed at this point.

Estimates are r.v.'s

They are functions of i.i.d samples
(training sets)

Estimates such as $\hat{\beta}_0$ and $\hat{\beta}_1$
have distributions and statistical
properties such as mean &
variance



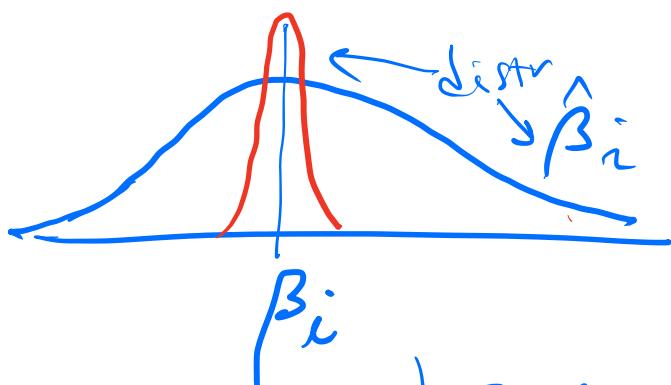
We wish $E[\hat{\beta}_i] = \beta_i$

i.e. $\hat{\beta}_i$ is unbiased

(Weaker property: asymptotic
unbiasedness)

$$\text{li } E[\hat{\beta}_i] = \beta_i$$

$n \rightarrow \infty$



Smaller $\text{Var}(\hat{\beta}_i)$
is more desirable
because we can rely
on the estimate

Efficiency:

$\hat{\beta}_i$ is called efficient

if

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\beta}_i) = 0$$

The standard deviation of
an estimate is called
its standard error in Statistics

Assessing the Accuracy of the Coefficient Estimates

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

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

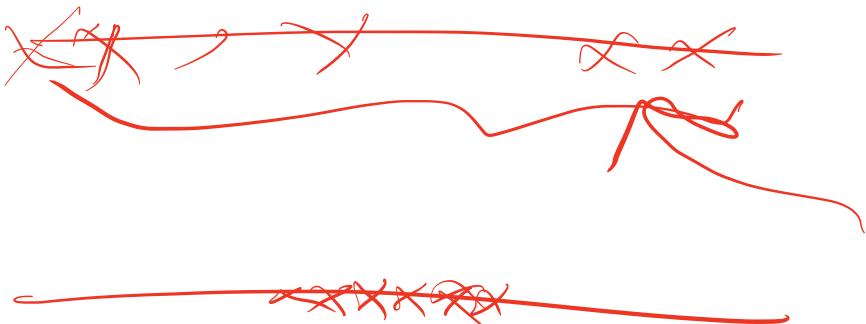
If σ^2 is large (more noise), we have more variability.

$\text{Var}(\hat{\beta}_1)$ where $\sigma^2 = \text{Var}(\varepsilon)$ in $\hat{\beta}_1$

If we have a sample of x_i 's

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)S_x^2$$

If we have "richer data,"
we have lower variability
in β_i



Normality of ε

- Note that in the following, the statistical results including confidence intervals and hypothesis testing assume that ε is normally distributed with mean zero and standard deviation σ .

$$\varepsilon \sim N(0, \sigma^2)$$

$$f_{\varepsilon}(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}$$

Assessing the Accuracy of the Coefficient Estimates

- These standard errors can be used to compute **confidence intervals**. A 95% confidence interval is defined as a range of values such that 95% of times, the range will contain the true unknown value of the parameter. It has the form $\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$.

*approximate
Interval Estimate*

Confidence intervals — continued

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

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

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

using fixed x_i 's

Confidence intervals — continued

In fact, an interval that will contain the true unknown value of the parameter β_1 in $1-\alpha$ percent of times is

$$[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)] \quad \alpha_2 = 0.05$$

Approximate CI for
 $1-\alpha=0.95$ (by the
textbook)

$1-\alpha$: level of confidence
 α : level of doubt

$$[\hat{\beta}_1 - t_{n-2,\alpha/2} \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,\alpha/2} \cdot \text{SE}(\hat{\beta}_1)]$$

More accurate CI

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot \text{SE}(\hat{\beta}_1)$$

Sample size

Student's t distribution

Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown.

$$X_i \sim N(\mu, \sigma^2)$$

i.i.d Sample of

$$\frac{\bar{X} - \mu}{S_x / \sqrt{n}} \sim t_{n-1}$$

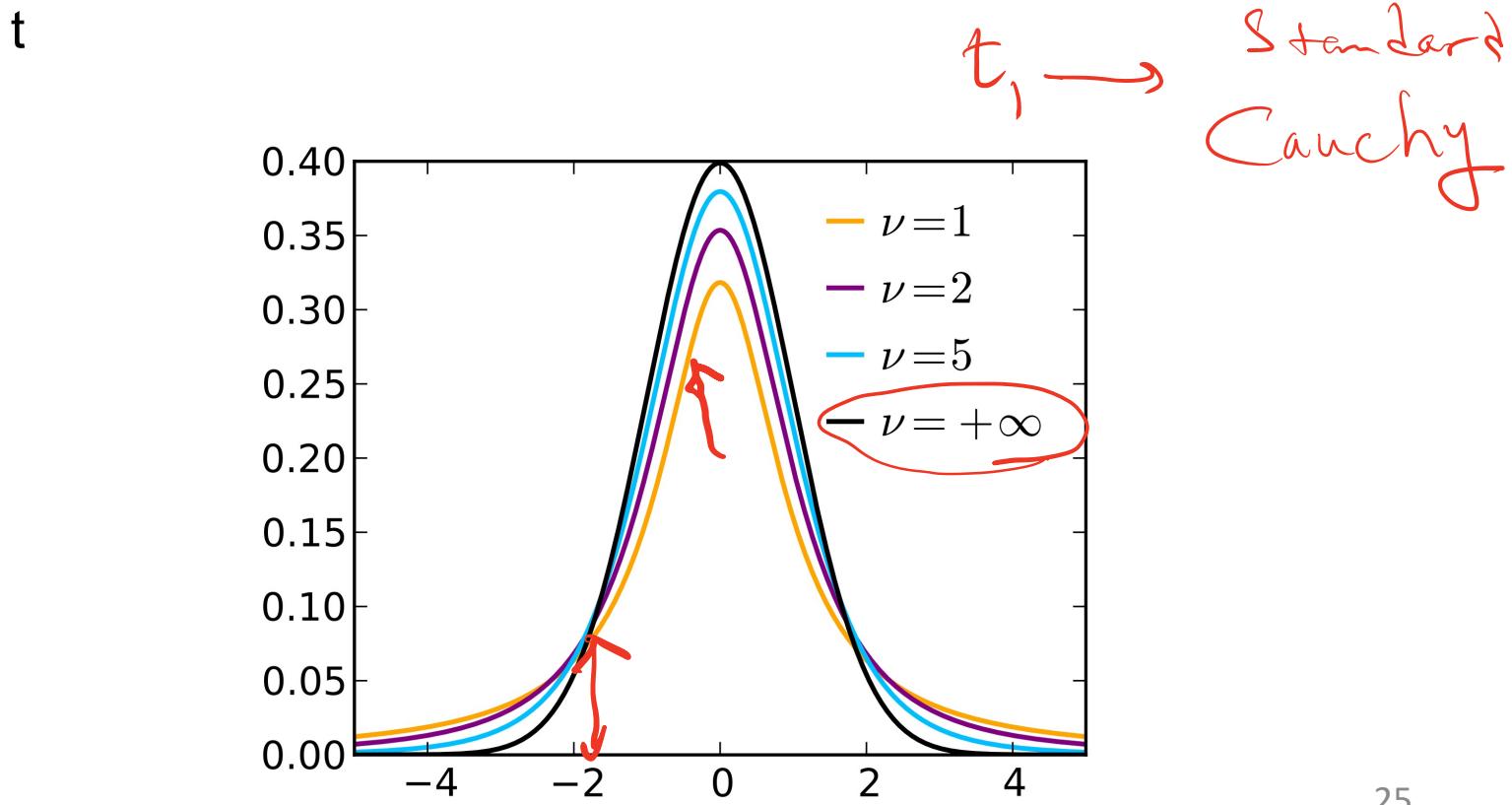
Student's t distribution

It was developed by William Sealy Gosset under the pseudonym Student.

The family is parameterized by a parameter ν , which is called the degrees of freedom.

The distribution is bell-shaped and has a zero mean, but its tails are heavier than the standard normal distribution.

Student's t distribution



Student's t distribution

When $\nu \rightarrow \infty$, $t_\nu \rightarrow Z$, where Z is a standard normal distribution, i.e., a normal distribution with mean zero and standard deviation 1.

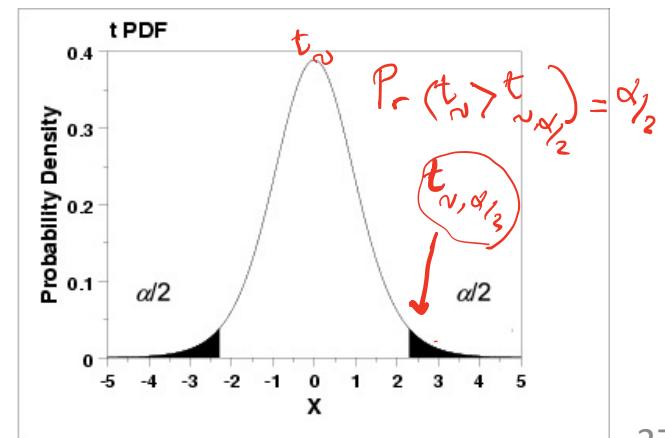
$$\begin{aligned} Z &\sim N(0, 1) \\ f_Z(z) &= \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \end{aligned}$$

Student's t distribution-cut off points

By $t_{n-2, \alpha/2}$, we mean:

$$\Pr(t_{n-2} > t_{n-2, \alpha/2}) = \alpha/2$$

In other words, the area under the pdf of the t distribution with $n-2$ degrees of freedom is $\alpha/2$ to the right of $t_{n-2, \alpha/2}$.



Advertisement Data for simple linear regression

```
lm.fit=lm(Sales~TV,data=Advertising) ## to get Table 3.1  
summary(lm.fit)
```

```
names(lm.fit)  
Call:  
lm(formula = sales ~ TV, data = Advertising)
```

```
coef(lm.fit)
```

```
confint(lm.fit)  
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 ***
TV	0.047537	0.002691	17.67	<2e-16 ***

```
---
```

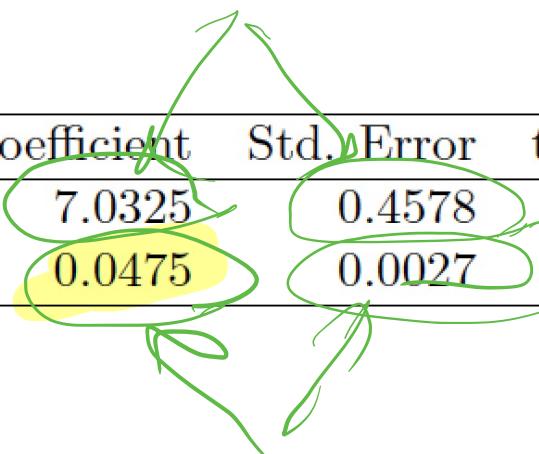
```
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
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
```

Statmodels

Results for 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



Confidence intervals — continued

For the advertising data, the 95% confidence interval for β_1 is approximately [0.042, 0.053]

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

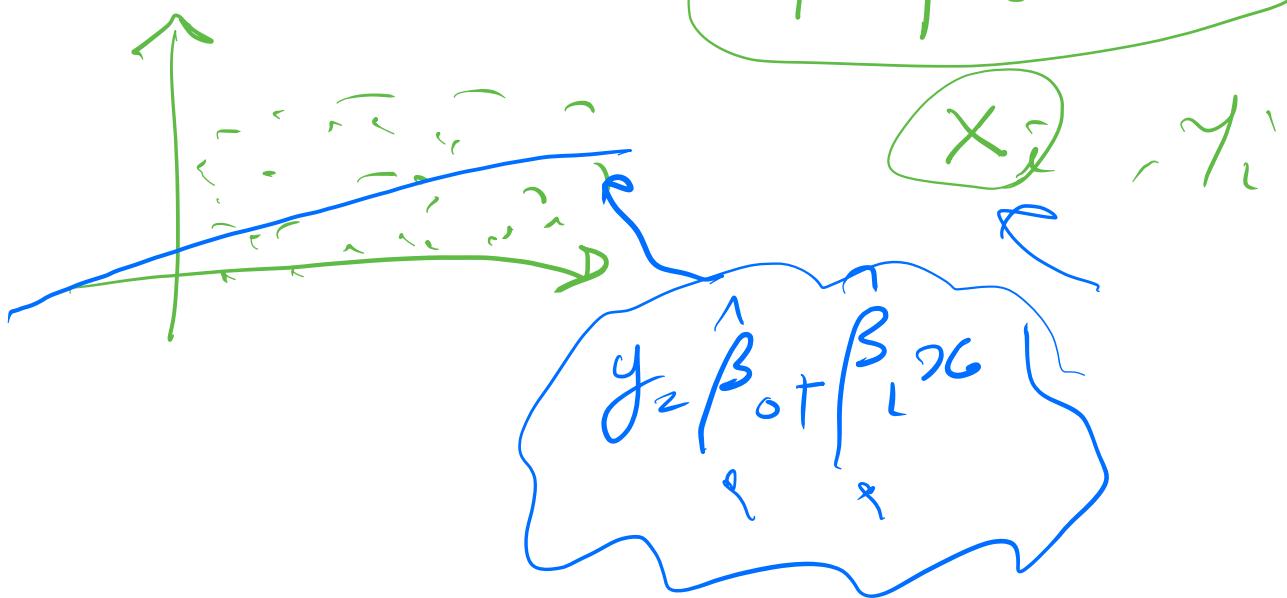
Approximate CI for
 $1-\alpha=0.95$ (by the
textbook)

$$[\hat{\beta}_1 - t_{n-2,\alpha/2} \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,\alpha/2} \cdot \text{SE}(\hat{\beta}_1)]$$

More accurate CI

If I give you data
sampled from

$$\beta_1 = 0$$
$$y = \beta_0 + \epsilon$$



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

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

versus the *alternative hypothesis*

$\{ H_A$:There is some relationship between X and Y.

$$Y = \beta_0 + \beta_1 X + \epsilon \quad \beta_1 \neq 0$$

Hypothesis testing

- Mathematically, this corresponds to testing

$$H_0: \beta_1 = 0 \quad Y_2 \beta_0 + \varepsilon$$

versus

$$H_A: \beta_1 \neq 0, \quad Y_2 \beta_0 + \beta_1 X + \varepsilon$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and X is not associated with Y .

For hypothesis testing

1. Assume a correct null hypothesis (innocent until proven guilty)

2. Choose a level of doubt (significance) α , which is the probability of rejecting a true null (convicting an innocent person)



Type-I error

$\alpha = \text{Prob}(\text{Type I error})$

$= \Pr_b$ (Reject H_0 & H_0 is true)

3. Assume the null hypothesis is TRUE
(under the null hypothesis)
 \equiv assuming H_0 is true

4. Present evidence and calculate the probability of
evidence being correct assuming
 H_0 is true

H_0 is true and if it is smaller than α , reject the null

Hypothesis testing — continued

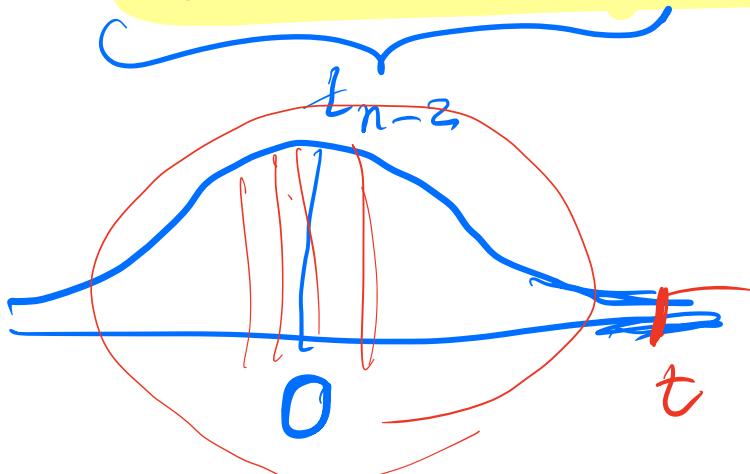
- To test the null hypothesis, we compute a *t-statistic*, given by

$$\beta_1 \neq 0$$

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

hypothesized β_1

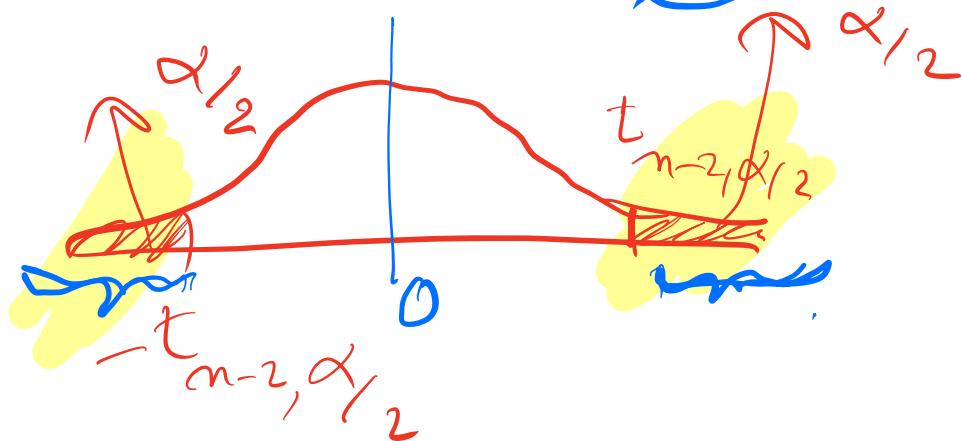
- This will have a *t-distribution* with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.



Under the null

Hypothesis testing — continued

- If the null hypothesis is true, the probability of observing $t > t_{n-2,\alpha/2}$ or $t < -t_{n-2,\alpha/2}$ would be α . α is the probability of rejecting a true null hypothesis, i.e. a **Type-I error**, and should be set **ahead of time** (metaphorically, by your boss). **Why?** Usually, α is selected to be 5%.



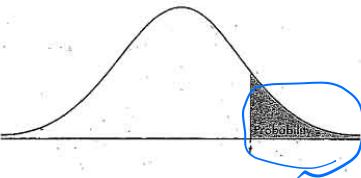
t_2 $\hat{\beta}_1 \}$
 $SE(\hat{\beta}_1) \}$

Small

→ rely on nonzero

$\hat{\beta}_1$

Rejection Region Approach



Sample size = 17
 $\alpha = 2\%$

TABLE B: t -DISTRIBUTION CRITICAL VALUES

df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.966	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.203	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.604	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.899	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.728	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.383	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.402	2.676	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.666	2.915	3.224	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

$$\begin{aligned} t_{m-2, \alpha/2} &= t_{15, 0.01} \\ &= 2.602 \end{aligned}$$

Rejection Region Approach

Reject the null $H_0: \beta_1 = 0$
if

$$t > t_{n-2, \alpha/2}$$

OR

$$t < -t_{n-2, \alpha/2}$$

otherwise, you fail to reject
the null.

Rejection Region Approach

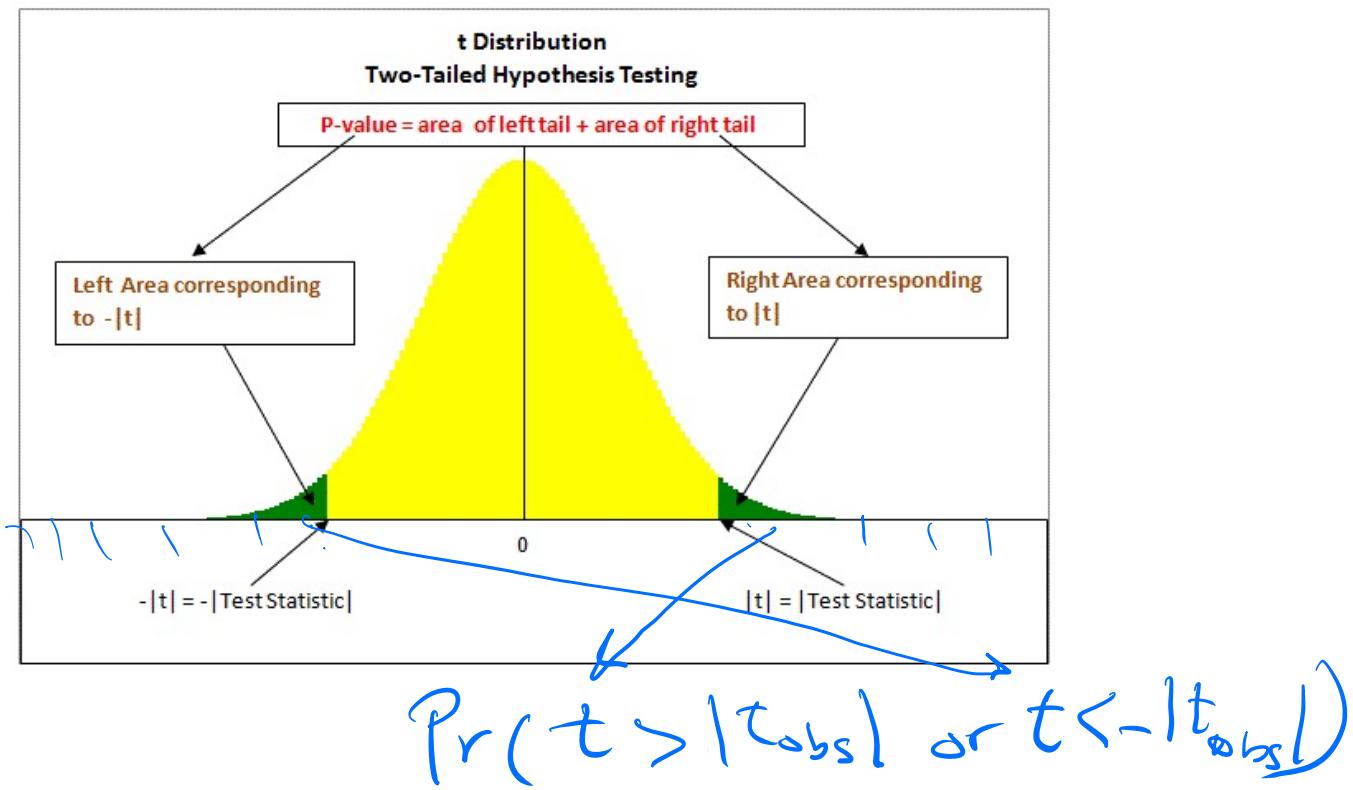
Hypothesis testing — continued

- 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*. *= Predictive*

$$t_{\text{obs}} = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

Hypothesis testing — continued

- We call this probability the *p-value*.



This p-value will be a measure of the quality of your evidence against H_0 .

~~Smaller p-values are better~~

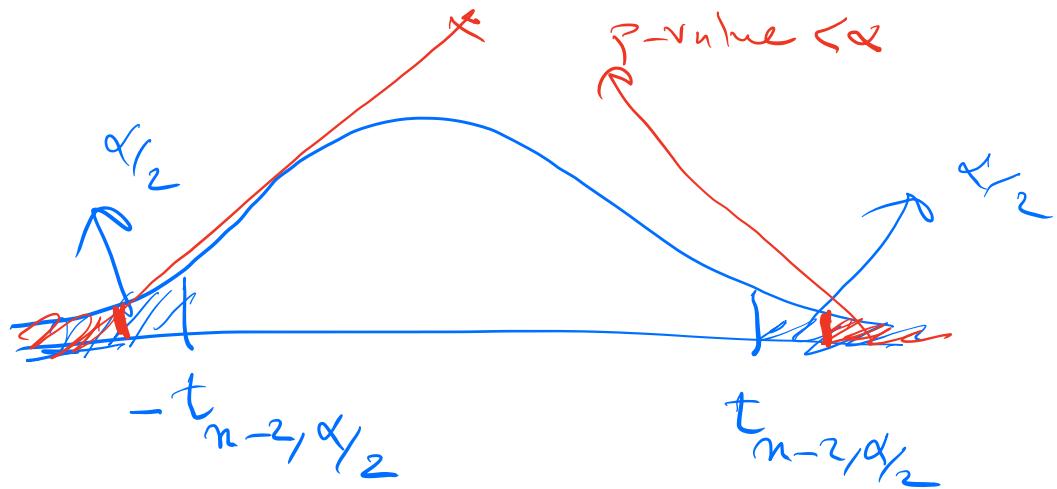
Reject the null if

$$p\text{-value} < \alpha$$

Hypothesis testing — continued

- If the p-value is very small, it means that the probability of seeing a t statistic extremer than what was observed assuming that $\beta_1 = 0$ is very small. So we reject the null.

Reject the null if $p < \alpha$



$p\text{-value} < \alpha \iff t_{\text{obs}}$ is
 in the rejection region

Advertisement Data for simple linear regression

```
lm.fit=lm(Sales~TV,data=Advertising) ## to get Table 3.1  
summary(lm.fit)
```

```
names(lm.fit)   Call:  
                lm(formula = sales ~ TV, data = Advertising)
```

```
coef(lm.fit)    Residuals:
```

```
confint(lm.fit) Min      1Q   Median      3Q      Max  
-8.3860 -1.9545 -0.1913  2.0671  7.2124
```

Coefficients:

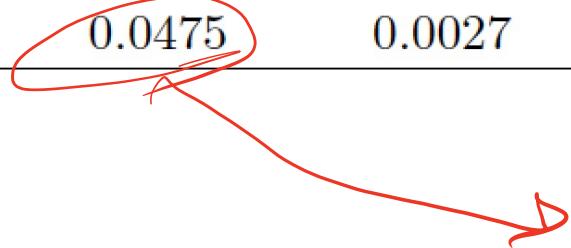
	Estimate	Std. Error	t value	Pr(> t)	pValue
(Intercept)	7.032594	0.457843	15.36	<2e-16	***
TV	0.047537	0.002691	17.67	<2e-16	***

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

Results for 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



Statistically
Significant
Predictor

Inferences about the Slope: t Test Example

Test Statistic: $t = 17.76$

From Software output:

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

$\hat{\beta}_1$

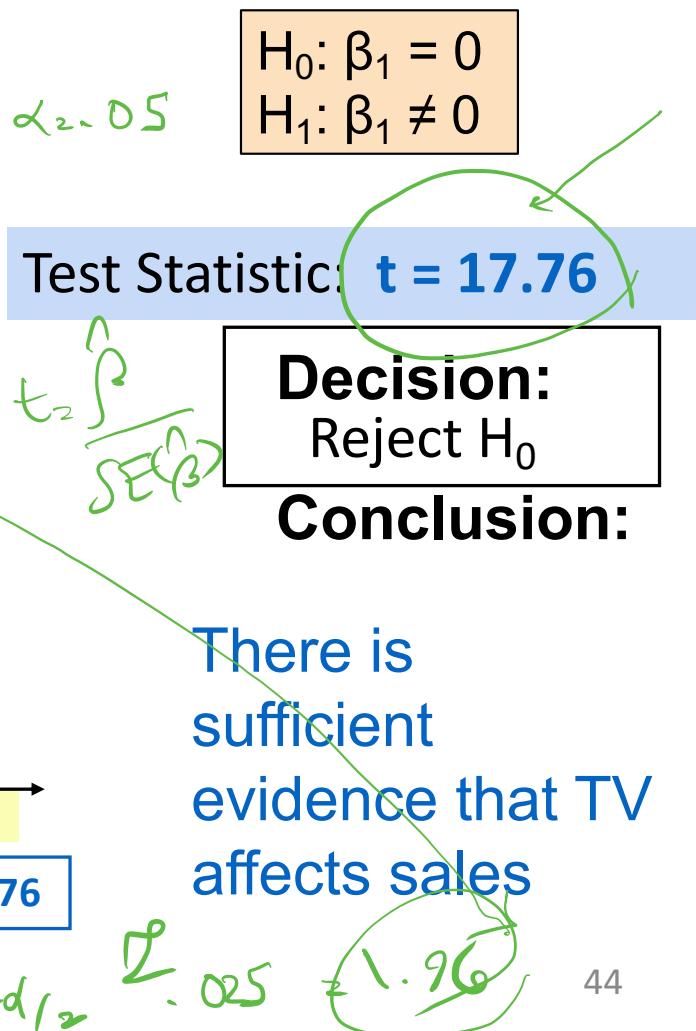
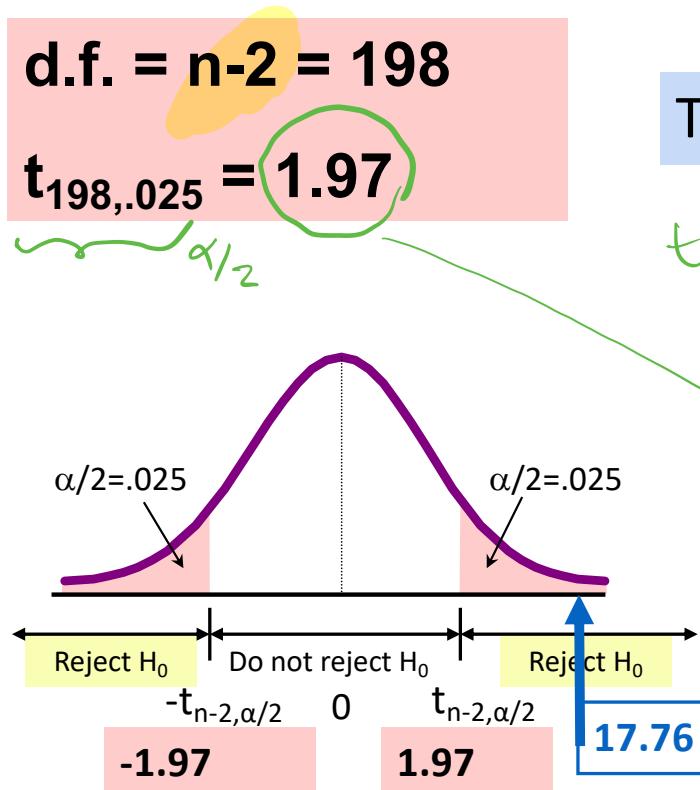
$\text{SE}(\hat{\beta}_1)$

t

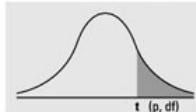
	Coefficients	Standard Error	t Stat	P-value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	.0475	0.0027	17.67	<0.0001

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned}$$

Inferences about the Slope: t Test Example



Numbers in each row of the table are values on a t -distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p).



Confidence:

$n=14 \quad 98\%$
 $t_{n-2, \alpha_2} = t_{12, 0.01}$
 $= 2.268$

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	—	—	80%	90%	95%	98%	99%	99.9%

Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

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

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

Assessing the Overall Accuracy of the Model

- The *Residual Standard Error* is used to estimate the variance of the noise ε , i.e. to measure how much on average the response deviated from the regression line.

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

Explanatory Power of a Linear Regression Equation

Total variation is made up of two parts:

$$\text{TSS} = \text{Regression SS} + \text{RSS}$$

Total Sum of Squares

Regression Sum of Squares

Error (residual) Sum of Squares

$$= \sum (y_i - \bar{y})^2$$

$$= \sum (\hat{y}_i - \bar{y})^2$$

$$= \sum (y_i - \hat{y}_i)^2$$

h-

where:

\bar{y} = Average value of the dependent variable

y_i = Observed values of the dependent variable

\hat{y}_i = Predicted value of y for the given x_i value

Explanatory Power of a Linear Regression Equation

TSS = total sum of squares

Measures the variation of the y_i values around their mean, \bar{y}

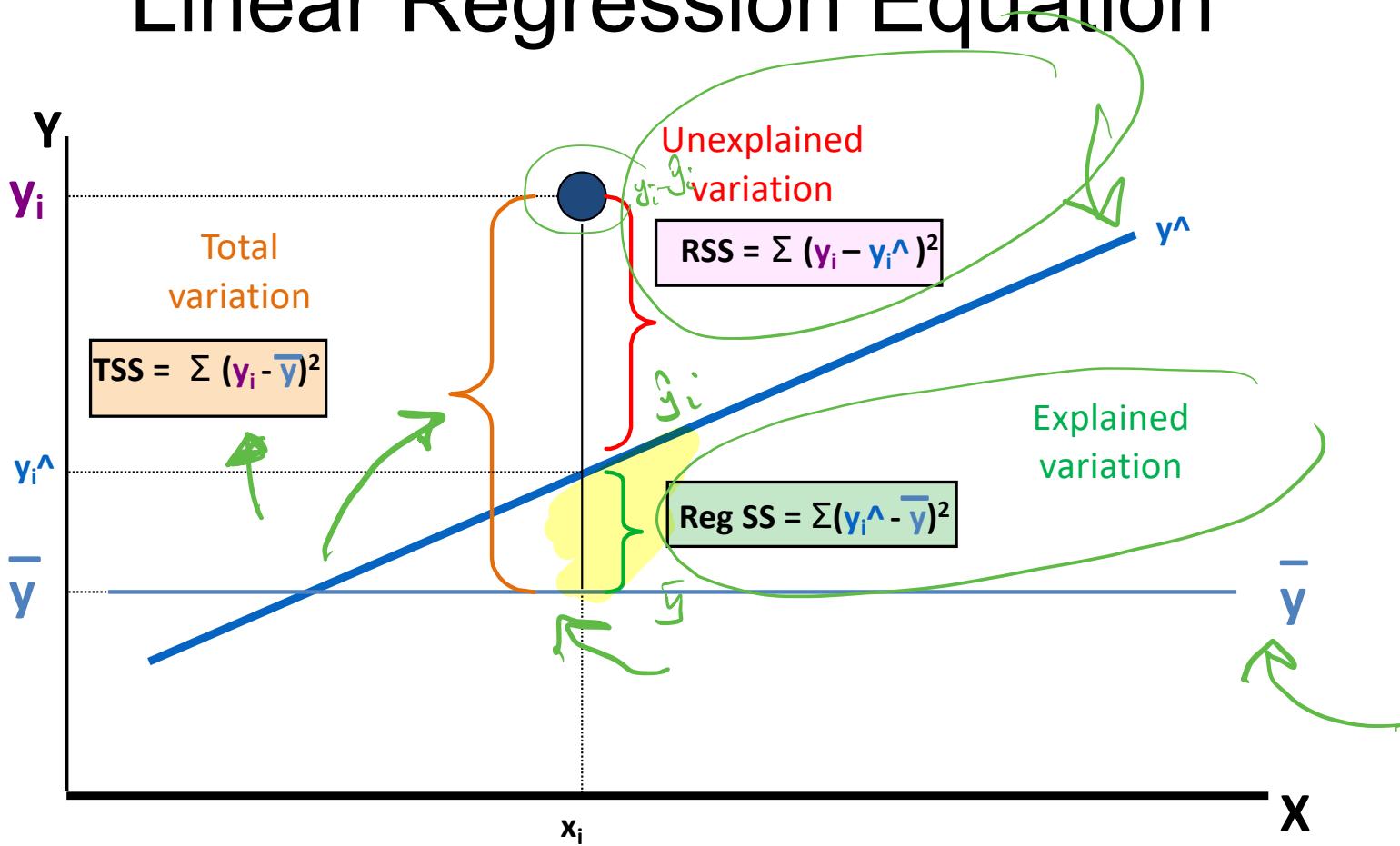
Regression SS = regression sum of squares

Explained variation attributable to the linear relationship between X and Y

RSS = Residual (error) sum of squares

Variation attributable to factors other than the linear relationship between X and Y

Explanatory Power of a Linear Regression Equation



Assessing the Overall Accuracy of the Model

- We are interested in the ratio of variation explained to total variation, i.e.

$$\frac{\text{RegSS}}{\text{TSS}} = \frac{\text{Explained}}{\text{Total}}$$

Handwritten annotations:

- RegSS is labeled "Explained"
- TSS is labeled "Total"
- $\text{TSS} - \text{RSS}$ is labeled "Explained"
- TSS is labeled "Total"

Assessing the Overall Accuracy of the Model

- *R-squared* or fraction of total variation 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*.

Can be computed for both
train & test data

Assessing the Overall Accuracy of the Model

- 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 :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$
$$= \frac{s_{XY}}{s_X s_Y}$$
$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

- We interpret β_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

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon.$$

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 β_j change in Y on average, while all the other variables stay fixed*”, are possible.

Interpreting regression coefficients

- 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.

Interpreting regression coefficients

- *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 β_j estimates the expected change in Y per unit change in X_j , *with all other predictors held fixed*. But predictors usually change together!

The woes of (interpreting) regression coefficients

- Example: Y total amount of change in your pocket;
 X_1 = # of coins; X_2 = # 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?

The woes of (interpreting) regression coefficients

- Y = number of tackles by a football player in a season; W and H are his weight and height.
- Fitted regression model is $\hat{Y} = b_0 + 0.50W - 0.10H$. How do we interpret $\hat{\beta}_2 < 0$?

Two quotes by famous Statisticians

“Essentially, all models are wrong, but some are useful”

George Box

Ptolemy

Two quotes by famous Statisticians

“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 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ as the values that minimize the sum of squared residuals RSS

Estimation and Prediction for Multiple Regression

- By defining:

$$\beta = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$$

$$y = [y_1 \ \dots \ \dots \ y_n]^T$$

Augmented
data / feature
matrix

$$X_{n \times (p+1)} = \begin{bmatrix} x_0 & x_1 & \dots & x_p \\ 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

$$x_0 = 1$$

Estimation and Prediction for Multiple Regression

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}}_{\hat{y}_i})^2.\end{aligned}$$

We can write RSS as:

$$\begin{aligned}RSS(\beta) &= \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)\end{aligned}$$

Estimation and Prediction for Multiple Regression

We can write RSS as:

$$\begin{aligned} RSS(\beta) &= \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \end{aligned}$$

Because:

$$\begin{aligned} \underline{\mathbf{y}} - \mathbf{X}\beta &= \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & \cdots & x_{1P} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nP} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_P \end{bmatrix} \\ &= \begin{bmatrix} y_1 - \beta_0 - \beta_1 x_{11} - \cdots - \beta_P x_{1P} \\ y_2 - \beta_0 - \beta_1 x_{21} - \cdots - \beta_P x_{2P} \\ \vdots \\ y_n - \beta_0 - \beta_1 x_{n1} - \cdots - \beta_P x_{nP} \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} \end{aligned}$$

Estimation and Prediction for Multiple Regression

Because:

$$\text{RSS}(\beta) = \|\underline{\epsilon}\|_2^2 = e_1^2 + \dots + e_n^2$$

x_i s and y_i s are known
(Supervised learning)

Estimation and Prediction for Multiple Regression

This is a quadratic function of β . So,

$$\frac{\partial}{\partial \beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) = \text{ } \delta$$

$$\frac{\partial}{\partial \beta} (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\beta - \beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X}\beta) = 0$$

$-2 \mathbf{y}^T \mathbf{X}\beta = 0$

$$\frac{\partial}{\partial c} (\underline{c}^T \mathbf{Q} \underline{c} + \underline{d}^T \underline{c} + b_0) = 2 \mathbf{Q} \underline{c} + \underline{d}$$

Given $\mathbf{Q} = \mathbf{Q}^T$

$$\frac{\partial}{\partial c} (cq_c + lc + b_0) = 2qc + l$$

scalar version

$$\cancel{\Phi} = -2 \cancel{X^T y} + 2 \cancel{X^T X \beta} = 0$$

$$X^T X \beta = X^T y$$

Normal Equation

If $X^T X$ is invertible

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

$$(X^T X)_{(p+1)}$$

X must be
full column rank

and this means the
features must be linearly
independent

Estimation and Prediction for Multiple Regression

Assuming that \mathbf{X} has full column rank, $\mathbf{X}^T\mathbf{X}$ is invertible, so the solution is

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



Estimation and Prediction for Multiple Regression

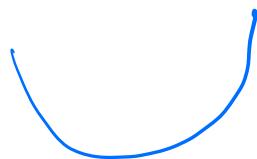
Also

$$\frac{\partial^2}{\partial \beta \partial \beta^T} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) = 2(\mathbf{X}^T \mathbf{X})$$

Hessian matrix

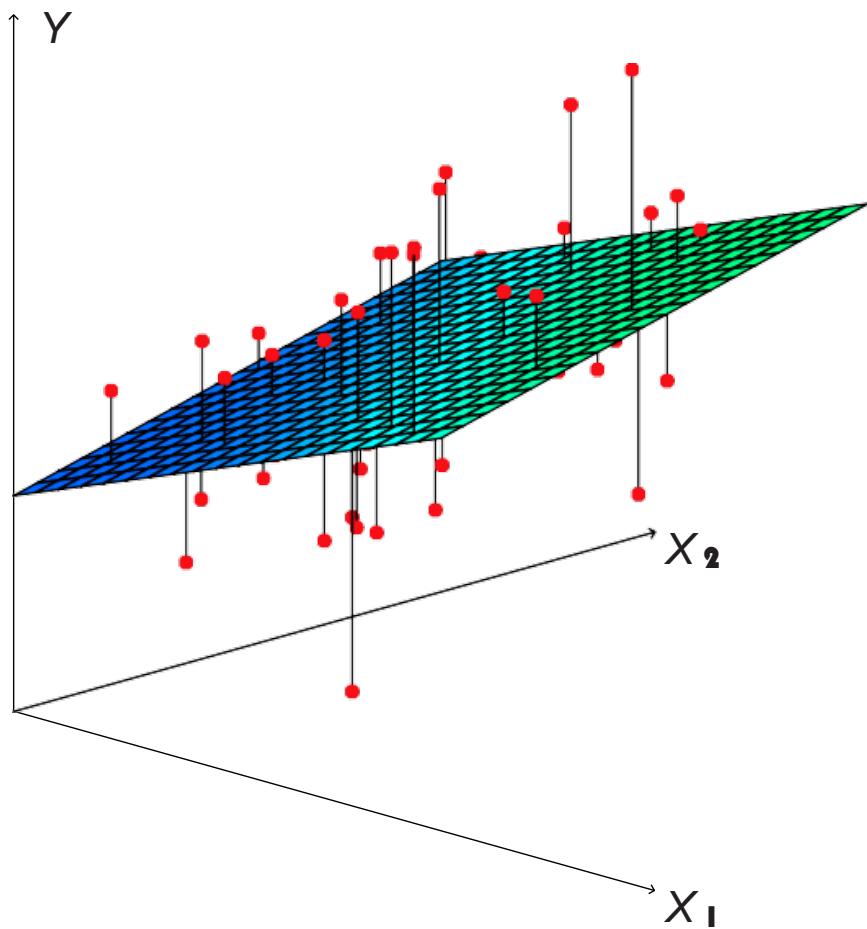
It is positive semi-definite

& $\hat{\beta}$ is
a minimizer



Estimation and Prediction for Multiple Regression

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



Estimation and Prediction for Multiple Regression

$$\begin{bmatrix} 1 & x_1' & \dots & x_p' \end{bmatrix}$$

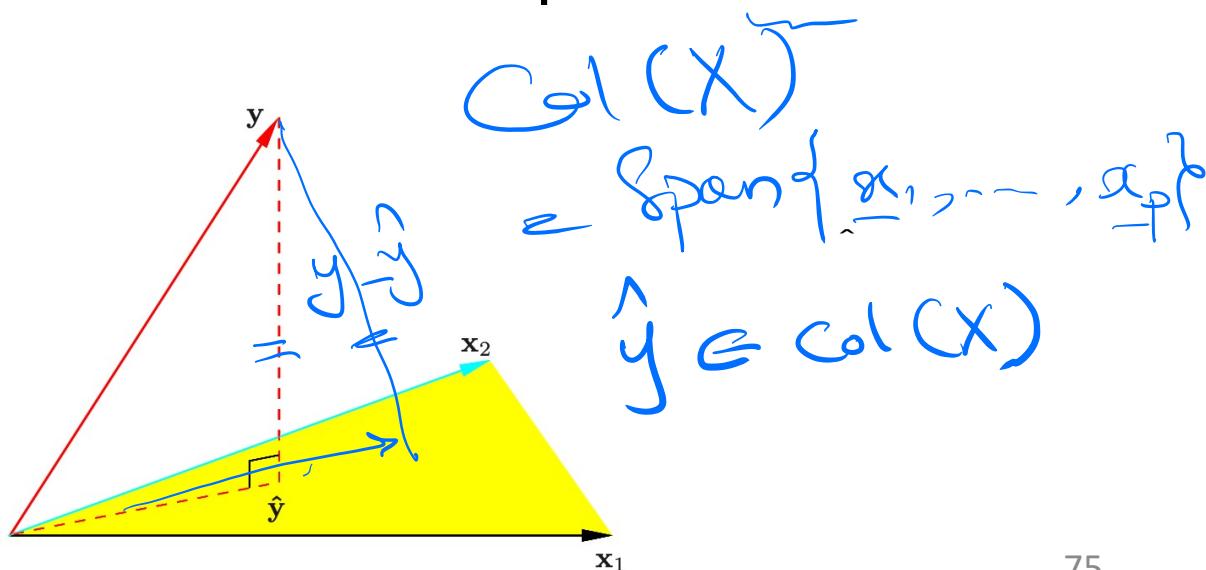
The predicted values at an input row vector \mathbf{x}' are given by $f(\mathbf{x}') = \underbrace{[1 \ \mathbf{x}']}_{\beta}^T \beta$; the fitted values at the training inputs are

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

smoothing

Orthogonality Principle

Theorem: $RSS(\beta) = \|y - X\beta\|^2$ is minimized if and only if the residual error $y - \hat{y}$ is orthogonal to the column space of X .



$y - \hat{y} \perp \text{Col}(X) \Rightarrow \|y - \hat{y}\|$ is minimum

Assume that $\tilde{y} \in \text{Col}(X)$ $\tilde{y} = X\tilde{\beta}$

We show that if $y - \hat{y} \perp \text{Col}(X)$,
then $\|y - \tilde{y}\|^2 \geq \|y - \hat{y}\|^2$

$$\begin{aligned}
 \|y - \tilde{y}\|^2 &= \|y - \hat{y} + \hat{y} - \tilde{y}\|^2 \\
 &= \underbrace{\langle y - X\hat{\beta} + X\hat{\beta} - X\tilde{\beta}, y - X\hat{\beta} + X\hat{\beta} - X\tilde{\beta} \rangle}_{\text{di}} \\
 &= \underbrace{\langle y - X\hat{\beta}, y - X\hat{\beta} \rangle}_{+} + \underbrace{\langle y - X\hat{\beta}, X\hat{\beta} - X\tilde{\beta} \rangle}_{+} \\
 &\quad + \underbrace{\langle X\hat{\beta} - X\tilde{\beta}, y - X\hat{\beta} \rangle}_{+} + \underbrace{\langle X\hat{\beta} - X\tilde{\beta}, X\hat{\beta} - X\tilde{\beta} \rangle}_{+} \\
 &= \|y - X\hat{\beta}\|^2 + 2\langle y - X\hat{\beta}, X\hat{\beta} - X\tilde{\beta} \rangle \\
 &\quad + \|X\hat{\beta} - X\tilde{\beta}\|^2 \quad \text{O} \quad X\tilde{\beta} \in \text{Col}(X)
 \end{aligned}$$

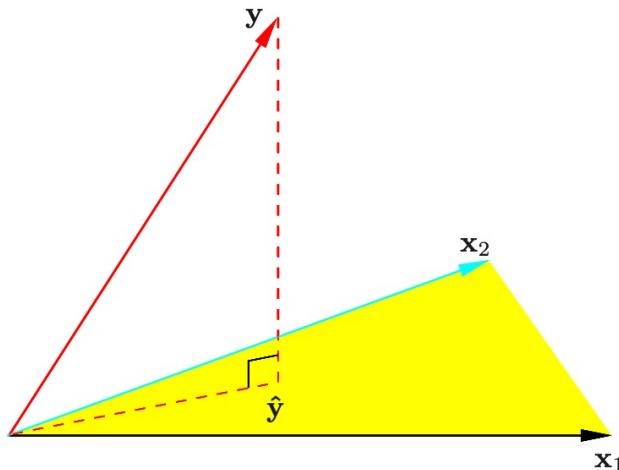
$$\hat{\beta} - \tilde{\beta} = \beta'$$

$$= \|y - x\hat{\beta}\|^2 + \underbrace{\|x\beta'\|_2^2}_{\geq 0}$$

$$\|y - \tilde{y}\|^2 \geq \|y - \underbrace{\hat{y}}_{x\beta}\|^2$$

Projection Matrix

The matrix $H = X(X^T X)^{-1} X^T$ is the projection matrix onto the space spanned by $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p$.



Orthogonal Projection Matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

- Symmetric: $\mathbf{H}^T = \mathbf{H}$
- Idempotent: $\mathbf{H}^2 = \mathbf{H}$
- $\mathbf{X}^T \mathbf{H} = \mathbf{X}^T$
- $\mathbf{X}^T (\mathbf{I} - \mathbf{H}) = \mathbf{0}$

the matrix of
projection onto
the subspace
orthogonal to $C(\mathbf{x})$

Orthogonal Projection

- Hence, \hat{y} is the orthogonal projection of y onto the column space of X

Dependent Columns

- If the columns of the matrix \mathbf{X} are linearly dependent (i.e. it is column rank deficient), there is an infinite number of solutions to the normal equation

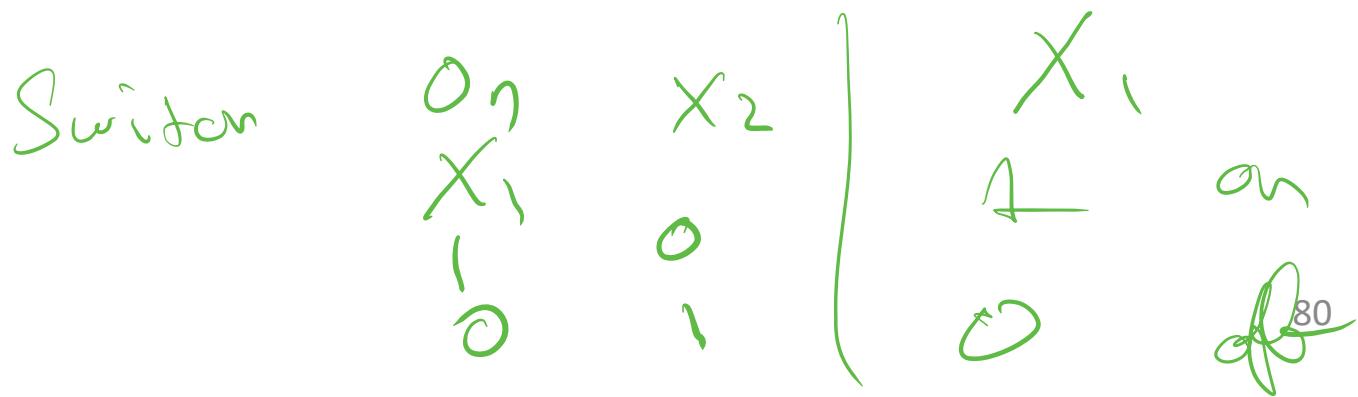
$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

- However, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is orthogonal projection of \mathbf{y} on the column space of \mathbf{X} for any solution $\hat{\beta}$ of the normal equation

There is a minimum norm
 $\hat{\beta}$ (Homework)

Dependent Columns

- This would occur, for example, if two of the inputs were perfectly correlated.
- The non-full-rank case also occurs when one or more qualitative inputs are coded in a redundant fashion.



Dependent Columns

- There is usually a natural way to resolve the non-unique representation, by recoding and/or dropping redundant columns in \mathbf{X} (more on this later).

Covariance Matrix of The Estimate

- The (variance-) covariance matrix of $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]^T\}$$
$$= \underbrace{\sigma^2}_{\text{green}} (\mathbf{X}^T \mathbf{X})^{-1} \quad \underbrace{[\hat{\beta} - E(\hat{\beta})]}_{\text{green}}$$

- This is because:

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

$$\underline{y} = \mathbf{X} \underline{\beta} + \underline{e}$$
$$E[\underline{y}] = \mathbf{X} \underline{\beta}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + e_i$$
$$\underline{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Covariance Matrix of The Estimate

$$\begin{aligned}\text{Var}(\hat{\beta}) &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]^T\} \\ &= \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}\end{aligned}$$

- This is because:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta + \mathbf{e}) \\ E[\hat{\beta}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta + 0 \\ &\sim \beta \quad \hat{\beta} \text{ is unbiased}\end{aligned}$$

$$\begin{aligned}
 \text{Var}(\hat{\beta}) &= E[(\hat{\beta} - E[\hat{\beta}])^T (\hat{\beta} - E[\hat{\beta}])] \\
 &= E[(\hat{\beta} - (\beta + (X^T X)^{-1} X^T e))^T (\hat{\beta} - (\beta + (X^T X)^{-1} X^T e))] \\
 &= E[(\hat{\beta} - (\beta + (X^T X)^{-1} X^T e))^T ((X^T X)^{-1} X^T e)] \\
 &= E[((X^T X)^{-1} X^T e)^T e] \\
 &= E[(X^T X)^{-1} X^T e e^T X (X^T X)^{-1}]
 \end{aligned}$$

ϵ iid $E[\epsilon_i \epsilon_j] = 0$

$$\left. \begin{aligned}
 E[\epsilon_i^2] &= \sigma^2 \\
 E[\epsilon_i] &= 0
 \end{aligned} \right\}$$

$$\begin{aligned}
 \hat{\beta}_2 &= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\
 &\approx \sigma^2 (X^T X)^{-1}
 \end{aligned}$$

Pseudo-inverse Learning

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

The matrix $\mathbf{X}^+ = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is called the pseudo-inverse of \mathbf{X} .

Learning $\hat{\beta}$ is a form of pseudo-inverse learning, which is a *one-pass* learning algorithm.

$$\mathbf{X}^+ \mathbf{X} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} = \mathbf{I}$$

Properties of Pseudo inverse (Moore-Penrose)

X^+ is a Pseudo-inverse of
 X iff

$$\checkmark (X^+ X)^T = X^+ X \text{ Symmetric}$$

$$\checkmark (X X^+)^T = X X^+$$

$$\checkmark X^+ X X^+ = X^+$$

$$\checkmark X X^+ X = X$$

$$X X^+ X^+ = X^+ \text{ commutes with } X$$

Unbiased Estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

RSS

- This means:

$$E[\hat{\sigma}^2] = \sigma^2$$

(Try to prove it?)

Normality of ε

- Note that in the following, the statistical results including confidence intervals and hypothesis testing assume that ε is normally distributed with mean zero and standard deviation σ .

$$\varepsilon \sim N(0, \sigma^2)$$

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

Distribution of The Estimates

$$\hat{\beta} \sim N(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

- This is a multivariate normal distribution with mean vector and variance-covariance matrix as shown.
- This is because:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \mathbf{y} = \mathbf{X}\beta + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

e_i 's are i.i.d normal

$\Rightarrow \mathbf{y}$ is jointly normal

$\rightarrow \beta$ is a linear transformation

of a jointly normal vector

$\rightarrow \beta$ is jointly normal

Distribution of The Estimates

$$\frac{\text{RSS}}{\sigma^2} = \frac{n - p - 1}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-p-1}^2$$

$\frac{1}{n-p-1}$ RSS

This is a chi-squared distribution with $n - p - 1$ degrees of freedom.

In addition, $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

Chi-Squared Distribution

- The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.
- The chi-squared distribution is a special case of the gamma distribution.

$Z_i \sim N(0,1)$, Z_i are i.i.d

$$Z_1^2 + \dots + Z_k^2 \sim \chi_k^2$$

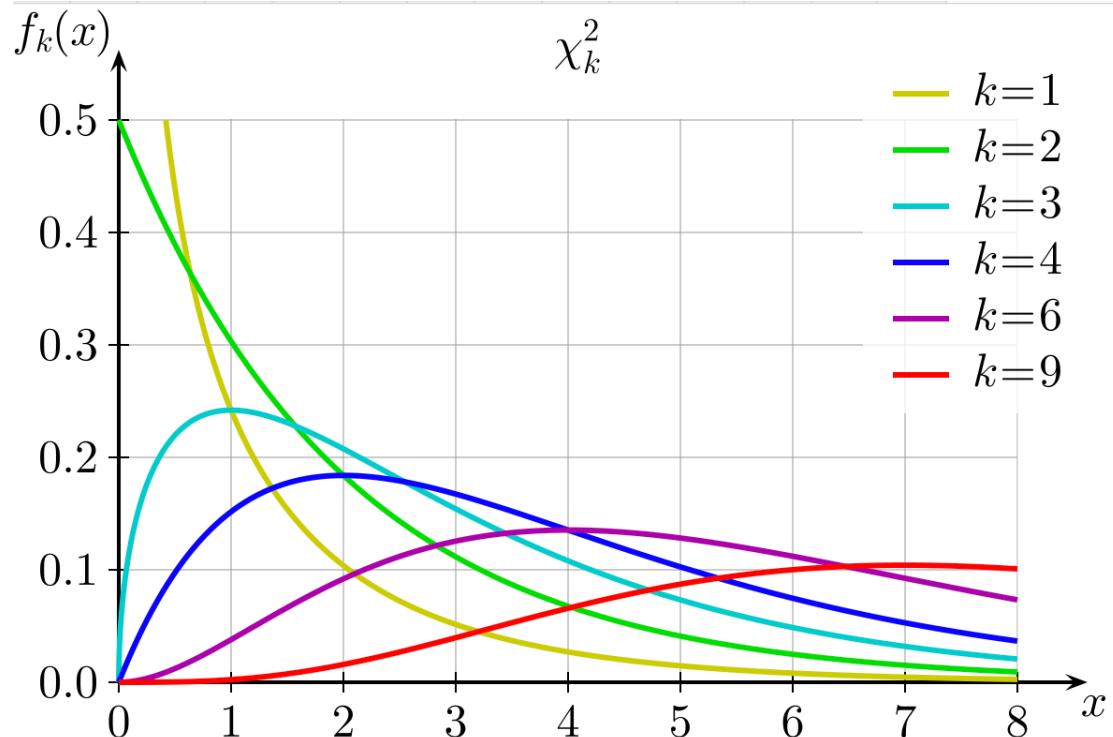
Chi-Squared Distribution

The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-squared distribution is a special case of the gamma distribution.

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}$$

Gamma(ρ, λ)
↓
 $\frac{k}{2}, \frac{1}{2}$

Chi-Squared Distribution



Student Distribution Again

If the true i^{th} parameter is β_i , then

$$\frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$$

where $\text{SE}(\hat{\beta}_i) = \sigma \sqrt{v_i}$ and v_i is the i^{th} diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$.

Var($\hat{\beta}_i$) is the i^{th} diagonal element of $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ Var($\hat{\beta}$)

Confidence intervals for Multiple Regression

An interval that will contain the true unknown value of the parameter β_i in $1-\alpha$ percent of times is

$$[\hat{\beta}_i - t_{n-p-1, \alpha/2} \cdot \text{SE}(\hat{\beta}_i), \hat{\beta}_i + t_{n-p-1, \alpha/2} \cdot \text{SE}(\hat{\beta}_i)]$$

$$\hat{\beta}_i \pm t_{n-p-1, \alpha/2} \text{SE}(\hat{\beta}_i)$$

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_i and Y .

versus the *alternative hypothesis*

H_A :There is some relationship between X_i and Y .

Hypothesis testing

- Mathematically, this corresponds to testing

$$H_0: \beta_i = 0$$

versus

$$H_A: \beta_i \neq 0,$$

since if $\beta_i = 0$ then X_i is not associated with Y .

Hypothesis testing

- In general, to test the following hypothesis

$$H_0: \beta_i = \beta$$

versus

$$H_A: \beta_i \neq \beta,$$

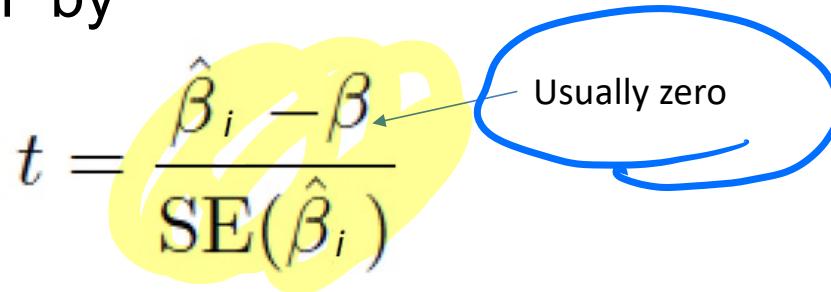
we use a t-statistic:

Hypothesis testing — continued

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

$$t = \frac{\hat{\beta}_i - \beta}{\text{SE}(\hat{\beta}_i)}$$

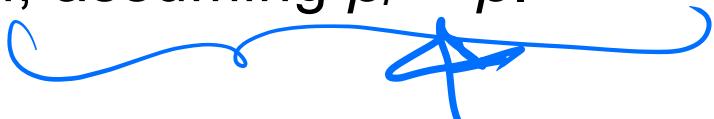
Usually zero



- This will have a *t*-distribution with $n - p - 1$ degrees of freedom, assuming $\beta_i = \beta$.

$$t = \frac{\hat{\beta}_i - \beta}{\text{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$$

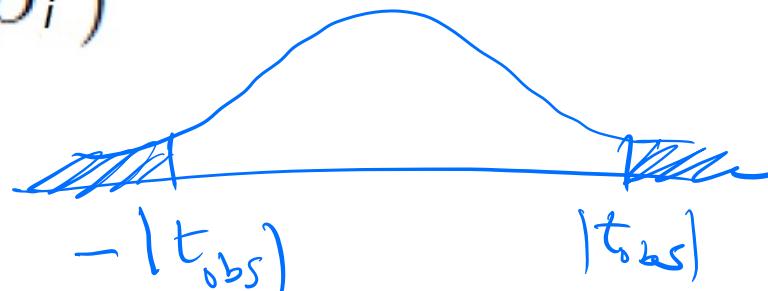
Under the null



Hypothesis testing — continued

- 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*.

$$t_{\text{obs}} = \frac{\hat{\beta}_i - \beta}{\text{SE}(\hat{\beta}_i)}$$



Hypothesis testing — continued

- If the p-value is very small, it means that the probability of seeing a t statistic extremer than what was observed (assuming that $\beta_i = \beta$) is very small. So we reject the null.

Rejection Region Approach

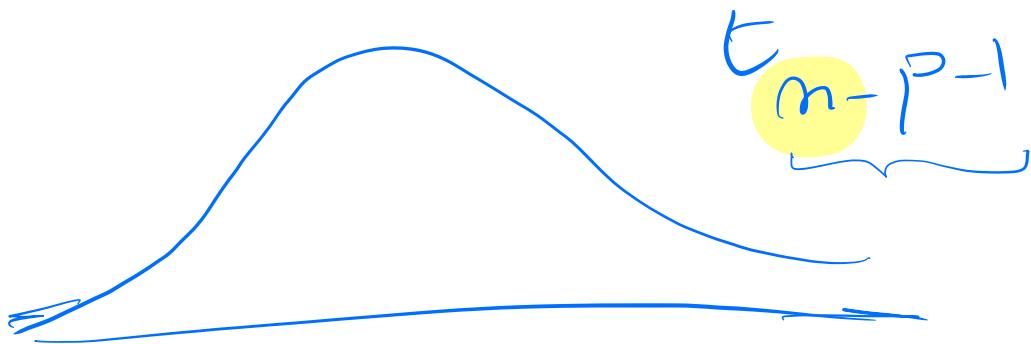
- Similar to simple regression

Reject the null if

$$t_{\text{obs}} > t_{n-p-1, \alpha/2}$$

OR

$$t < -t_{n-p-1, \alpha/2}$$



What happens if we have a large number of features, p large: to reject the null we need stronger evidence (larger $|t_{\text{obs}}|$)

Or larger $n \equiv$ more data

Results for advertising data

	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

Correlations:

	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

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?*

Some important questions

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?*

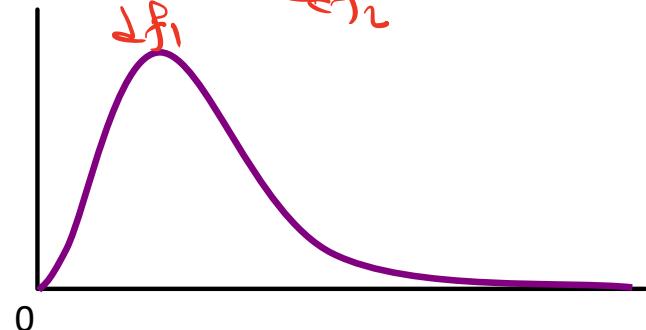
Is at least one predictor useful?

For the first question, we can use the F-statistic

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

null:
 $y_i = \beta_0 + \epsilon_i$
 $\epsilon_i \sim N(0, \sigma^2)$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570



Tests on Regression Coefficients

Tests on All Coefficients

F-Test for Overall Significance of the Model

Shows if there is a linear relationship between **all** of the X variables considered together and Y

Use F test statistic

Hypotheses:

$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ (no linear relationship)

$H_1: \text{at least one } \beta_i \neq 0$ (at least one independent variable affects Y)

F-Test for Overall Significance

Test statistic:

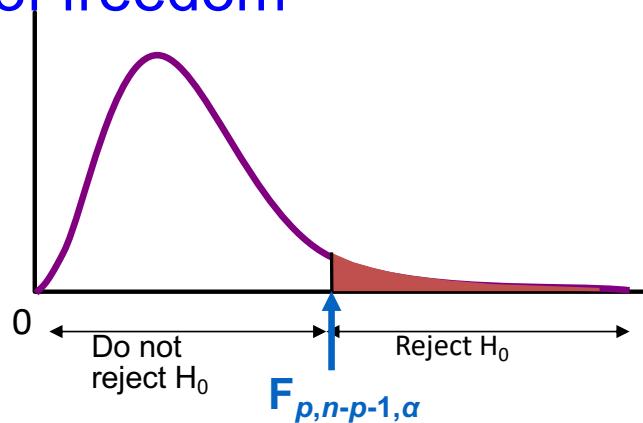
$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)} \sim F_{p, n-p-1}$$

where F has p (numerator) and $(n - p - 1)$ (denominator) degrees of freedom

The decision rule is

Reject H_0 if $F > F_{p, n-p-1, \alpha}$

$\nearrow \text{Not } \alpha/2$



$$\alpha_2 = 0.05 \quad n_2 = 16 \quad p_2 = 9 \quad F_{p_1, n-p_1, d} = F_{9, 6, .05}$$

F-Test for Overall Significance

$\approx 4.$

F - Distribution ($\alpha = 0.05$ in the Right Tail)

Denominator Degrees of Freedom	df ₂	Numerator Degrees of Freedom								
		1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988	
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	

F-Test for Overall Significance

F - Distribution ($\alpha = 0.01$ in the Right Tail)

Denominator Degrees of Freedom df ₂	df ₁	Numerator Degrees of Freedom								
		1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761	
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188	
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106	
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315	
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875	
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911	
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297	
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948	
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804	
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822	
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971	
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225	
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567	
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981	
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458	
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986	
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560	
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172	
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818	
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494	
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195	
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920	
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665	
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876	
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185	
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586	
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073	

F-Test for Overall Significance

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

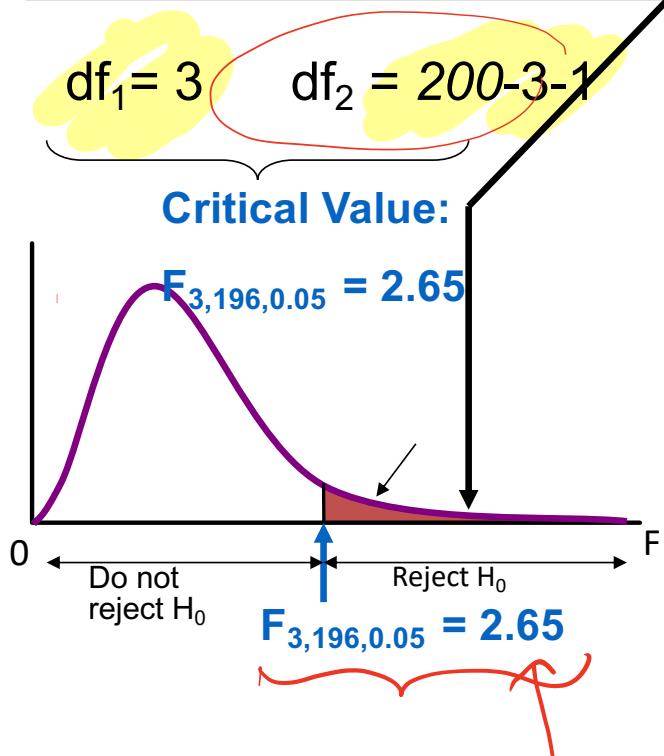
$H_1:$ Not all three of $\beta_1, \beta_2, \beta_3$ are zero

Test Statistic: $F=570$

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

Decision:

Since F test statistic is in the rejection region ($p\text{-value} < .05$), reject H_0



Conclusion:

There is evidence that at least one independent variable affects Y

The Gauss-Markov Theorem

One of the most famous results in statistics asserts that the least squares estimates of the parameters β have the *smallest variance* among all linear unbiased estimates.

BLUE (Best Linear
Unbiased Estimate)

The Gauss-Markov Theorem

Consider estimation of any linear combination of the parameters $\theta = \mathbf{a}^T \beta$

Example: $y = \underbrace{\mathbf{x}^T \beta}$ for some test point \mathbf{x} .

The least squares estimate of $\mathbf{a}^T \beta$

is

$$\hat{\theta} = \mathbf{a}^T \hat{\beta} = \underbrace{\mathbf{a}^T}_{\text{Column vector of } \mathbf{a}} \underbrace{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\text{Matrix } \mathbf{A}} \underbrace{\mathbf{y}}_{\text{Vector } \mathbf{y}}$$

(why?)

$$\begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \beta$$

The Gauss-Markov Theorem

Considering X to be fixed, this is a linear function $\mathbf{c}_0^T \mathbf{y}$ of the response vector \mathbf{y} .

If we assume that the linear model is correct, $\mathbf{a}^T \hat{\boldsymbol{\beta}}$ is unbiased since

The least squares estimate of any linear combination of $\boldsymbol{\beta}$'s is a linear function of the response vector \mathbf{y} .

$$\begin{aligned} \mathbb{E}[\underline{\alpha}^\top \hat{\beta}] &= \mathbb{E}[\underline{\alpha}^\top (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{y}] \\ &= \underline{\alpha}^\top (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underbrace{\mathbb{E}[\underline{y}]}_{\underline{X}\underline{\beta}} = \underline{\alpha}^\top \underline{\beta} \\ \underline{y} &\sim \underline{X}\underline{\beta} + \underline{e} \end{aligned}$$

The Gauss-Markov Theorem

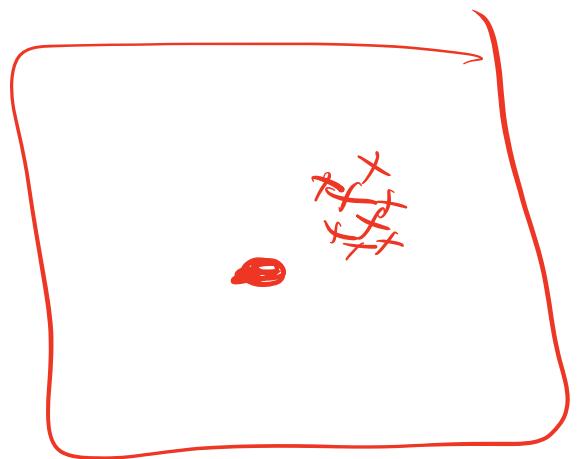
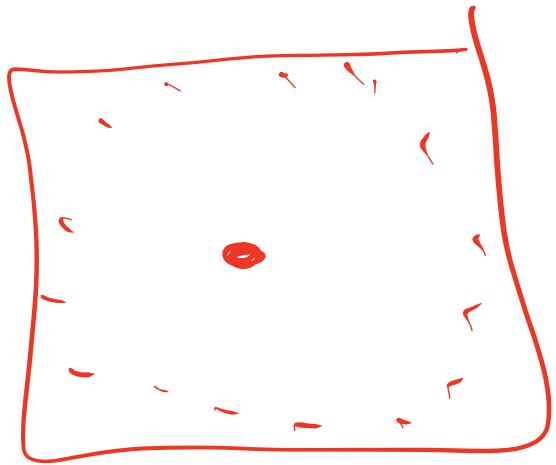
The Gauss–Markov theorem states that if we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ that is unbiased for $\mathbf{a}^T \beta$, i.e. $E[\mathbf{c}^T \mathbf{y}] = \mathbf{a}^T \beta$, we have

$$\text{Var}(a^T \hat{\beta}) \leq \text{Var}(\mathbf{c}^T \mathbf{y}).$$

$\hat{\theta} = \mathbf{a}^T \hat{\beta}$ is the minimum variance estimate among all linear unbiased estimates of θ

Gauss - Markov is

not saying that the least squares is the best estimate in terms of MSE even if the underlying model is linear, because there might be biased linear estimates that have lower variance and win in the bias-variance trade-off.



The Gauss-Markov Theorem

Consider the mean squared error of an estimator $\tilde{\theta}$ in estimating θ :

$$\begin{aligned}\text{MSE}(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 \\ &= \underbrace{\text{Var}(\tilde{\theta})}_{\text{A}} + \underbrace{[E(\tilde{\theta}) - \theta]^2}_{\text{B}}.\end{aligned}$$

The Gauss-Markov Theorem

$$\begin{aligned}\text{MSE}(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 \\ &= \text{Var}(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2.\end{aligned}$$

- The first term is the variance, while the second term is the squared bias.
- The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error of all linear estimators with no bias.

The Gauss-Markov Theorem

- There may well exist a **biased estimator** with smaller mean squared error.
- Such an estimator would trade a little bias for a larger reduction in variance.

The Gauss-Markov Theorem

- Biased estimates are commonly used. Any method that shrinks or sets to zero some of the least-squares coefficients may result in a biased estimate.
- We discuss many examples, including variable subset selection and ridge regression later.

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.

Variance

Bio

GREEDY

Deciding on the important variables

- However, we often can't examine all possible models, since they are 2^p of them; for example, when $p = 40$ there are over a trillion models!
- Instead, we need an automated approach that searches through a subset of them. We discuss two commonly used approaches next.

Forward selection

$$y = \beta_0 + \epsilon$$
$$\hat{\beta}_0 = \bar{y}$$

- 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.

Forward selection

- 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.

Backward selection

- 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.

Model selection — continued

- These include Mallow's C_p ,
Akaike information criterion
(AIC), Bayesian information
criterion (BIC), adjusted R^2 and
Cross-validation (CV).

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*.

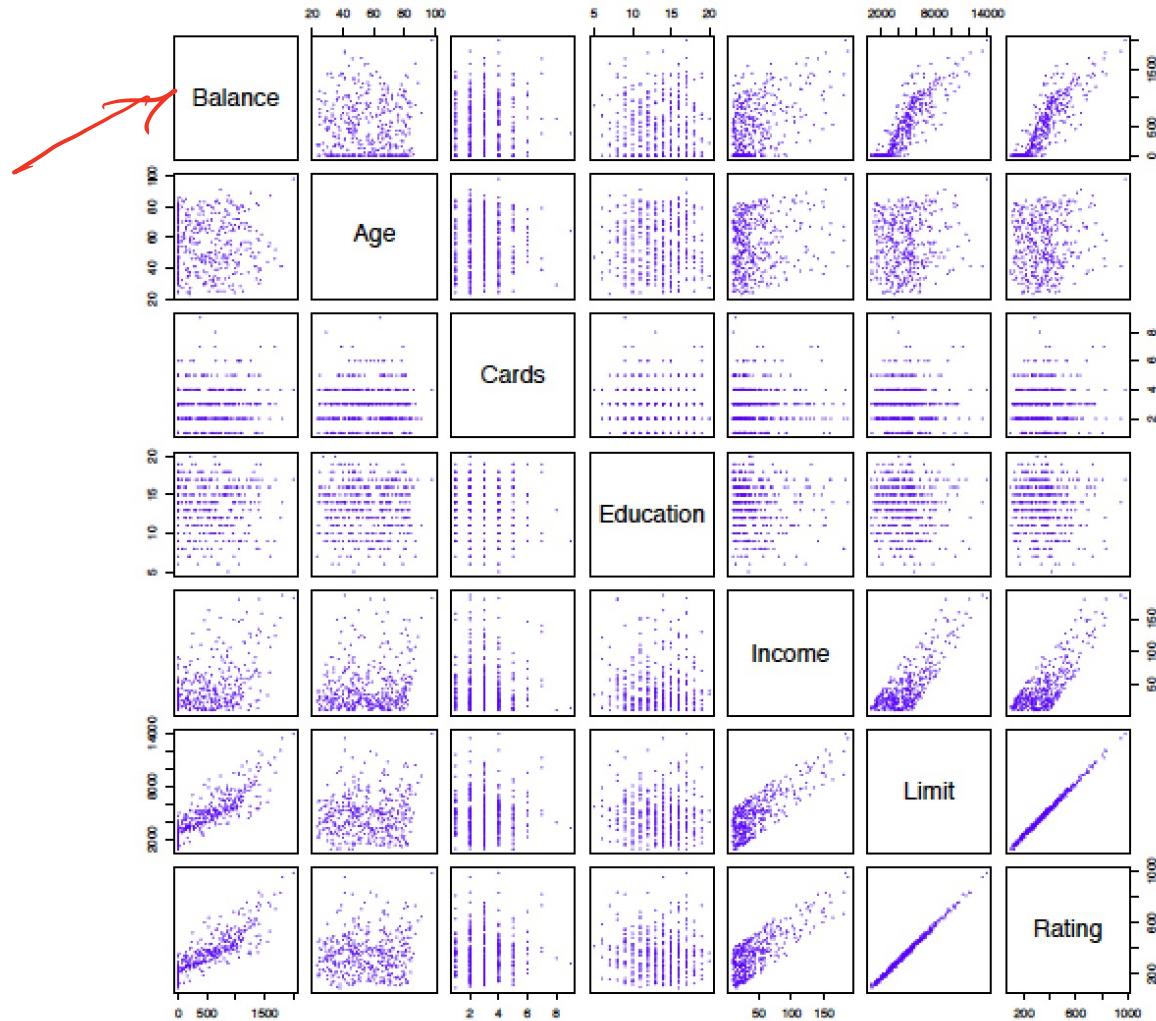


Other Considerations in the Regression Model

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).

Credit Card Data



Qualitative Predictors — cont'd

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable *dummy 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:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Intrepretation?

$$\begin{aligned} y_i^2 &= \beta_0 + \epsilon_i \\ \beta_0 &= \bar{y} \end{aligned}$$

Credit card data — continued

Results for gender model:

	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

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

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

The equation shows a regression model where the dependent variable y_i is predicted by intercept β_0 , qualitative predictors x_{i1} and x_{i2} , and error term ϵ_i . The coefficients β_1 and β_2 are crossed out with red wavy lines. The term $\beta_0 + \epsilon_i$ is highlighted with a yellow circle and has a red arrow pointing to the text 'if i th person is AA.'. The term $\beta_1 + \epsilon_i$ is associated with 'Asian' and the term $\beta_2 + \epsilon_i$ is associated with 'Caucasian'.

Qualitative predictors with more than two levels

- 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*.

Results for ethnicity

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity [Asian]	-18.69	65.02	-0.287	0.7740
ethnicity [Caucasian]	-12.50	56.68	-0.221	0.8260

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.

Extensions of the Linear Model

- For example, the linear model

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

states that the average effect on **sales** of a one-unit increase in **TV** is always β_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.

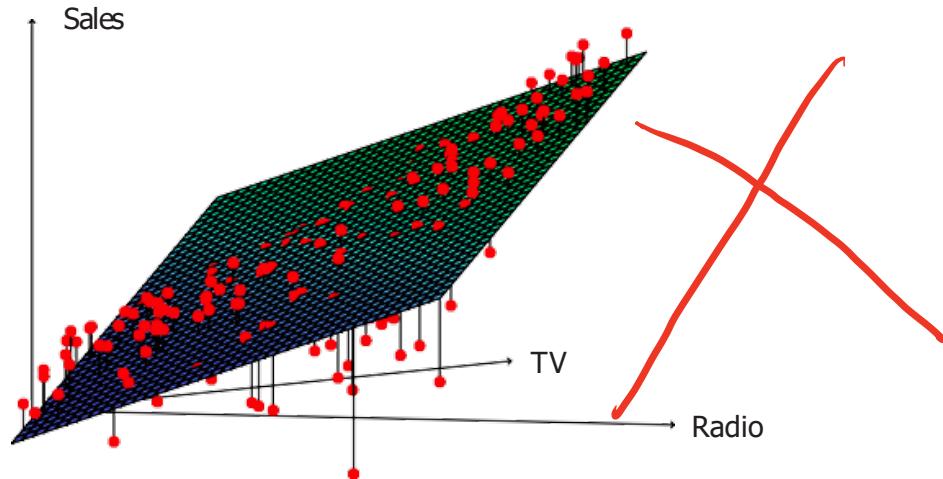
Interactions — continued

- 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**.

Interactions — continued

- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

Interaction in the Advertising data?



When levels of either **TV** or **radio** are low, then the true **sales** are lower than predicted by the linear model.

But when advertising is split between the two media, then the model tends to underestimate **sales**.

Modelling interactions — Advertising data

Model takes the form

$$\begin{aligned} \text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon \end{aligned}$$

β_0 β_1 β_2 β_3 $x_1 x_2$

Results:

	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×radio	0.0011	0.000	20.73	< 0.0001



Interpretation

- The results in this table suggest that interactions are important.
- The p-value for the interaction term $\text{TV} \times \text{radio}$ is extremely low, indicating that there is strong evidence for $H_A: \beta_3 \neq 0$.

Interpretation

- 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 $\frac{(96.8 - 89.7)}{(100 - 89.7)} = 69\%$ of the variability in **sales** that remains after fitting the additive model has been explained by the interaction term.

Interpretation — continued

- 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.

Interpretation — continued

- An increase in radio advertising of \$1, 000 will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$ 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 *hierarchical 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 — continued

- 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.

Interaction between Quantitative and Qualitative Variables

Consider the **Credit** data set, and suppose that we wish to predict **balance** using **income** (quantitative) and **student** (qualitative).

Without an interaction term, the model takes the form

$$\begin{aligned}\text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}\end{aligned}$$

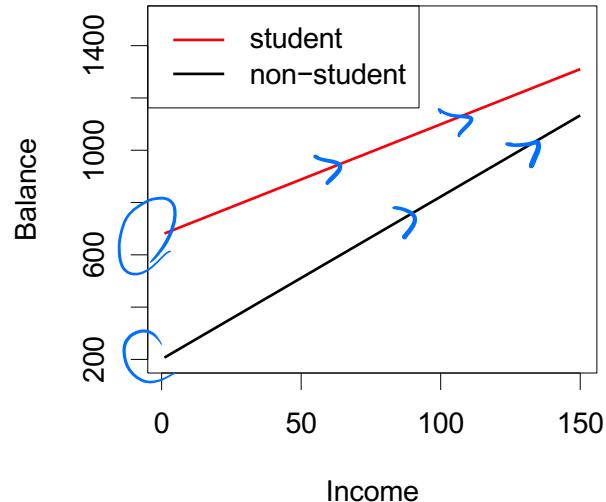
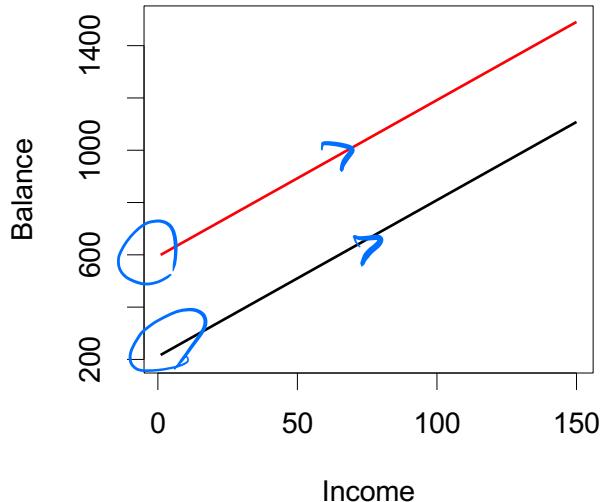
$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{student}$

Interaction between Quantitative and Qualitative Variables

With interactions, it takes the form

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

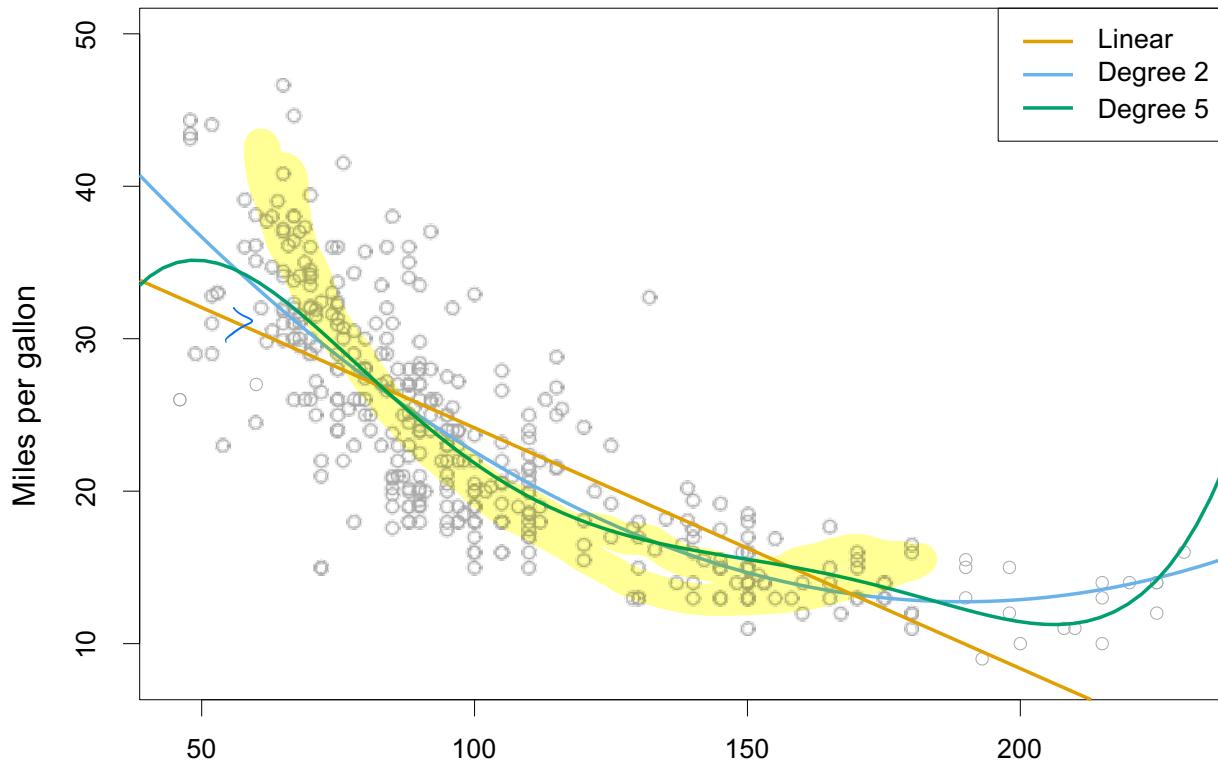
$$\text{balance}_i = \begin{cases} \beta_0 + \beta_1 \text{ income} + \beta_2 \text{ student} + \beta_3 \text{ income} \times \text{student} & \text{Student} \\ \beta_0 + \beta_1 \text{ income} & \text{Non-Student} \end{cases}$$



Credit data; 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



$$\beta_0 + \beta_1 X + \beta_2 X^2$$

Horsepower

The figure suggests that

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

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower ²	0.0012	0.0001	10.1	< 0.0001

A yellow highlight covers the first three rows of the table, corresponding to the terms in the regression equation. A blue arrow points from the handwritten note 'better fit' to the p-value of the horsepower² term.

F-Test for Significance of Groups of Coefficients

- We need to test for the significance of groups of coefficients simultaneously.
- **Example:** testing if a categorical variable with k levels can be excluded from a model, we need to test whether the coefficients of the dummy variables used to represent the levels can all be set to zero. Here we use the F statistic as well.

F-Test for Significance of Groups of Coefficients

- Here we use the F statistic,

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1 - 1)},$$

- RSS_1 : associated with the bigger model with $p_1 + 1$ parameters
- RSS_0 : associated with the nested smaller model with $p_0 + 1$ parameters, having $p_1 - p_0$ parameters constrained to be zero.

F-Test for Significance of Groups of Coefficients

The F statistic measures the change in residual sum-of-squares per additional parameter in the bigger model, and it is normalized by an estimate of σ^2

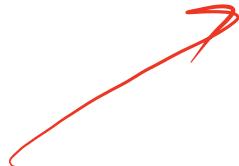
$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(n - p_1 - 1)},$$

Under the Gaussian assumptions, and the null hypothesis that the smaller model is correct, the F statistic will have a $F_{p_1-p_0, n-p_1-p_0}$ distribution.

What we did not cover

- ✓ Outliers
- ✓ Non-constant variance of error terms
- ✓ High leverage points
- ✓ Collinearity

See text Section 3.33



→ ISLR

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

Generalizations of the Linear Model

- *Classification problems:* logistic regression, support vector machines
- *Non-linearity:* kernel smoothing, splines and generalized additive models; nearest neighbor methods.

Generalizations of the Linear Model

- *Interactions*: Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- *Regularized fitting*: Ridge regression and lasso

Appendix: More on Qualitative/ Categorical Variables

Material from:

<https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/>

	male	X_1	X_2	Fence
F		0	1	
M		1	0	$X_2 = 1 - X_1$

	X_1	X_2	X_3
Green	Green	Grey	Brown
Green	1	0	0
Grey	0	1	0
Brown	0	0	1

$X_3 = 1 - X_1 - X_2$

The above method (one-hot encoding) creates one dummy variable for

each category - It
creates problems for
linear regression, but it's
usually safe to use
with other (nonlinear) ML
algorithms

Another way: binary
coders

{ $x_1 \ x_2 \ x_3 \ x_4 \ x_5$
{ 2^5 (or $2^5 - 1$) levels
using 5 bits

Qualitative/ Categorical Variables

- Challenges faced while dealing with categorical variables:
- A categorical variable has too many levels. This pulls down performance level of the model. For example, a cat. variable “zip code” would have numerous levels.

Qualitative/ Categorical Variables

- Challenges faced while dealing with categorical variables:
- A categorical variable has levels which rarely occur. Many of these levels have minimal chance of making a real impact on model fit. For example, a variable ‘disease’ might have some levels which would rarely occur.

Qualitative/ Categorical Variables

- Challenges faced while dealing with categorical variables:
- There is one level which always occurs i.e. for most of the observations in data set there is only one level. Variables with such levels fail to make a positive impact on model performance due to very low variation.

Qualitative/ Categorical Variables

- Challenges faced while dealing with categorical variables:
- If the categorical variable is masked, it becomes a laborious task to decipher its meaning. Such situations are commonly found in data science competitions.

Qualitative/ Categorical Variables

- Challenges faced while dealing with categorical variables:
- You can't fit categorical variables into a regression equation in their raw form. They must be treated.

Qualitative/ Categorical Variables

- Challenges faced while dealing with categorical variables:
- Most of the algorithms (or ML libraries) produce better result with numerical variable. In python, library “sklearn” requires features in numerical arrays.

Methods to deal with Qualitative/Categorical Variables

Convert to Number

Label Encoder: It is used to transform non-numerical labels to numerical labels (or nominal categorical variables). Numerical labels are always between 0 and n_classes-1.

Methods to deal with Qualitative/Categorical Variables

Label Encoder:

```
In [53]: train.head(5)
```

```
Out[53]:
```

	sex	pclass
0	male	3
1	female	1
2	female	3
3	female	1
4	male	3

```
In [54]: from sklearn.preprocessing import LabelEncoder
```

```
number = LabelEncoder()
train['sex'] = number.fit_transform(train['sex'].astype('str'))
test['sex'] = number.fit_transform(test['sex'].astype('str'))
```

```
train.head(5)
```

```
Out[54]:
```

	sex	pclass
0	1	3
1	0	1
2	0	3
3	0	1
4	1	3

Methods to deal with Qualitative/Categorical Variables

Label Encoder: A common challenge with nominal categorical variable is that, it may decrease performance of a model. For example: We have two features “age” (range: 0-80) and “city” (81 different levels).

Methods to deal with Qualitative/Categorical Variables

Now, when we'll apply label encoder to 'city' variable, it will represent 'city' with numeric values range from 0 to 80. The 'city' variable is now similar to 'age' variable since both will have similar data points, which is certainly not a right approach.

Methods to deal with Qualitative/ Categorical Variables

Convert to Number

Convert numeric bins to number: Let's say, bins of a continuous variable are available in the data set (shown next).

Methods to deal with Qualitative/Categorical Variables

Convert to Number

Convert numeric bins to number

User_ID	Product_ID	Gender	Age	Occupatio	City_Cate	Stay_In_C	Marital_St	Product_C	Product_C	Product_C	Purchase
1000001	P00069042	F	0-17	10 A		2	0	3			8370
1000001	P00248942	F	0-17	10 A		2	0	1	6	14	15200
1000001	P00087842	F	0-17	10 A		2	0	12			1422
1000001	P00085442	F	0-17	10 A		2	0	12	14		1057
1000002	P00285442	M	55+	16 C	4+		0	8			7969
1000003	P00193542	M	26-35	15 A		3	0	1	2		15227
1000004	P00184942	M	46-50	7 B		2	1	1	8	17	19215
1000004	P00346142	M	46-50	7 B		2	1	1	15		15854
1000004	P0097242	M	46-50	7 B		2	1	1	16		15686
1000005	P00274942	M	26-35	20 A		1	1	8			7871
1000005	P00251242	M	26-35	20 A		1	1	5	11		5254

Methods to deal with Qualitative/ Categorical Variables

Convert to Number

Convert numeric bins to number:

Variable “Age” has bins (0-17, 17-25, 26-35 ...). We can convert these bins into definite numbers using the following methods:

Using label encoder for conversion. But, these numerical bins will be treated same as multiple levels of non-numeric feature. Hence, wouldn’t provide any additional information

Methods to deal with Qualitative/ Categorical Variables

Convert to Number

Convert numeric bins to number:

Variable “Age” has bins (0-17, 17-25, 26-35 ...). We can convert these bins into definite numbers using the following methods:

Create a new feature using mean or mode (most relevant value) of each age bucket. It would comprise of additional weight for levels.

Methods to deal with Qualitative/Categorical Variables

Convert to Number

Convert numeric bins to number:

User_ID	Product_ID	Gender	Age	New_Age	Occupatio	City_Cate	Stay_In_C	Marital_St	Product_C	Product_C	Product_C	Purchase
1000001	P00069042	F	0-17	14	10	A		2	0	3		8370
1000001	P00248942	F	0-17	14	10	A		2	0	1	6	1422
1000001	P00087842	F	0-17	14	10	A		2	0	12		1422
1000001	P00085442	F	0-17	14	10	A		2	0	12	14	1057
1000002	P00285442	M	55+	60	16	C	4+		0	8		7969
1000003	P00193542	M	26-35	30	15	A		3	0	1	2	15227
1000004	P00184942	M	46-50	47	7	B		2	1	1	8	19215
1000004	P00346142	M	46-50	47	7	B		2	1	1	15	15854
1000004	P0097242	M	46-50	47	7	B		2	1	1	16	15686
1000005	P00274942	M	26-35	30	20	A		1	1	8		7871
1000005	P00251242	M	26-35	30	20	A		1	1	5	11	5254

Methods to deal with Qualitative/Categorical Variables

Convert to Number

Convert numeric bins to number:

Variable “Age” has bins (0-17, 17-25, 26-35 ...). We can convert these bins into definite numbers using the following methods:

Create two new features, one for lower bound of age and another for upper bound. In this method, we’ll obtain more information about these numerical bins compare to earlier two methods.

Methods to deal with Qualitative/Categorical Variables

Convert to Number

Convert numeric bins to number:

User_ID	Product_ID	Gender	Age	Lower_Age	Upper_Age	Occupatio	City_Cate	Stay_In_C	Marital_St	Product_C	Product_C	Product_C	Purchase
1000001	P00069042	F	0-17	0	17	10	A	2	0	3			8370
1000001	P00248942	F	0-17	0	17	10	A	2	0	1	6	14	15200
1000001	P00087842	F	0-17	0	17	10	A	2	0	12			1422
1000001	P00085442	F	0-17	0	17	10	A	2	0	12	14		1057
1000002	P00285442	M	55+	55	80	16	C	4+	0	8			7969
1000003	P00193542	M	26-35	26	35	15	A	3	0	1	2		15227
1000004	P00184942	M	46-50	46	50	7	B	2	1	1	8	17	19215
1000004	P00346142	M	46-50	46	50	7	B	2	1	1	15		15854
1000004	P0097242	M	46-50	46	50	7	B	2	1	1	16		15686
1000005	P00274942	M	26-35	26	35	20	A	1	1	8			7871
1000005	P00251242	M	26-35	26	35	20	A	1	1	5	11	.	5254

Methods to deal with Qualitative/ Categorical Variables

Combine Levels: one can sometimes simply combine the different levels. There are various methods of combining levels.

Here are commonly used ones:
Using Business Logic

Methods to deal with Qualitative/ Categorical Variables

Combine Levels: Using Business Logic

For example, we can combine levels of a variable “zip code” at state or district level. This will reduce the number of levels and improve the model performance also.

Methods to deal with Qualitative/ Categorical Variables

Combine Levels: Using Business Logic

Zip Code	District
110044	South Delhi
110048	South Delhi
110049	South Delhi
110006	North Delhi
110007	North Delhi
110058	West Delhi
110059	West Delhi
110063	West Delhi
110064	West Delhi

Methods to deal with Qualitative/ Categorical Variables

**Combine Levels:
Using frequency or response rate:**

When we don't have domain knowledge about the levels, we combine levels by considering the frequency distribution or response rate.

Methods to deal with Qualitative/ Categorical Variables

Combine Levels:

Using frequency or response rate:

Consider the frequency distribution of each level and combine levels having frequency less than 5% of total observation (5% is standard but you can change it based on distribution). This is an effective method to deal with rare levels.

Methods to deal with Qualitative/ Categorical Variables

Combine Levels:

Using frequency or response rate:

We can also combine levels by considering the response rate of each level. We can simply combine levels having similar response rate into same group.

Methods to deal with Qualitative/ Categorical Variables

Combine Levels:

Using frequency or response rate:

Finally, you can also look at both frequency and response rate to combine levels. You first combine levels based on response rate then combine rare levels to relevant group.

Methods to deal with Qualitative/Categorical Variables

Combine Levels:

Based on Frequency

Levels	Frequency	New_Level
HA001	9%	HA001
HA002	12%	HA002
HA003	4%	New
HA004	1%	New
HA005	3%	New
HA006	11%	HA006
HA007	1%	New
HA008	4%	New
HA009	10%	HA009
HA010	4%	New
HA011	8%	HA011
HA012	12%	HA012
HA013	3%	New
HA014	11%	HA014
HA015	2%	New
HA016	4%	New
HA017	0%	New

Based on Response Rate

Levels	Response_Rate	New_Level
HA014	98%	1
HA001	97%	1
HA003	93%	1
HA009	81%	2
HA015	75%	3
HA010	73%	3
HA006	66%	4
HA017	60%	4
HA007	49%	5
HA004	36%	6
HA005	31%	6
HA012	28%	7
HA008	25%	7
HA013	23%	7
HA016	22%	7
HA002	21%	8
HA011	5%	9

Based on Frequency and Response Rate

Levels	Frequency	Response_Rate	New_Level1	New_Level2
HA014	11%	98%	1	1
HA001	9%	97%	1	1
HA003	4%	93%	1	1
HA009	10%	81%	2	2
HA015	2%	75%	3	2
HA010	4%	73%	3	2
HA006	11%	66%	4	4
HA017	0%	60%	4	4
HA007	1%	49%	5	4
HA004	1%	36%	6	4
HA005	3%	31%	6	4
HA012	12%	28%	7	7
HA008	4%	25%	7	7
HA013	3%	23%	7	7
HA016	4%	22%	7	7
HA002	12%	21%	8	8
HA011	8%	5%	9	9

Methods to deal with Qualitative/Categorical Variables

Dummy Coding

Dummy coding is a commonly used method for converting a categorical input variable into continuous variable. ‘Dummy’, as the name suggests is a duplicate variable which represents one level of a categorical variable.

Methods to deal with Qualitative/Categorical Variables

Dummy Coding

Presence of a level is represented by 1 and absence is represented by 0. For every level present, one dummy variable will be created. Look at the representation below to convert a categorical variable using dummy variable.

Methods to deal with Qualitative/Categorical Variables

Dummy Coding

```
In [46]: train.head(5)
```

```
Out[46]:
```

	sex	pclass
0	male	3
1	female	1
2	female	3
3	female	1
4	male	3

```
In [47]: train=train=pd.get_dummies(train)  
train.head(5)
```

```
Out[47]:
```

	pclass	sex_female	sex_male
0	3	0	1
1	1	1	0
2	3	1	0
3	1	1	0
4	3	0	1

Methods to deal with Qualitative/ Categorical Variables

Dummy Coding

Note: Assume, we have 500 levels in categorical variables. Then, should we create 500 dummy variables? If you can automate it, very well. Or else, I'd suggest you to first, reduce the levels by using combining methods and then use dummy coding. This would save your time. This method is also known as “One_Hot Encoding”.

Methods to deal with Qualitative/ Categorical Variables

Feature Hashing

Read:

<https://blog.myyellowroad.com/using-categorical-data-in-machine-learning-with-python-from-dummy-variables-to-deep-category-66041f734512>