清华大学统计学辅修课程

Linear Regression Analysis

Lecture 8-
Multiple Linear Regression:
Example & Inference

周在莹

清华大学统计学研究中心

http://www.stat.tsinghua.edu.cn

清华大学统计学研究中心

Topic 1:
Multiple Linear Regression Example

Outline

- Description of the Example
- Descriptive Summaries
- Investigation of Various Models
- Conclusions

Study of CS Students

- Too many computer science majors at university dropping out of program
- Want to find predictors of success to be used in the admission process
- Predictors must be available at time of entry into program

Data Available

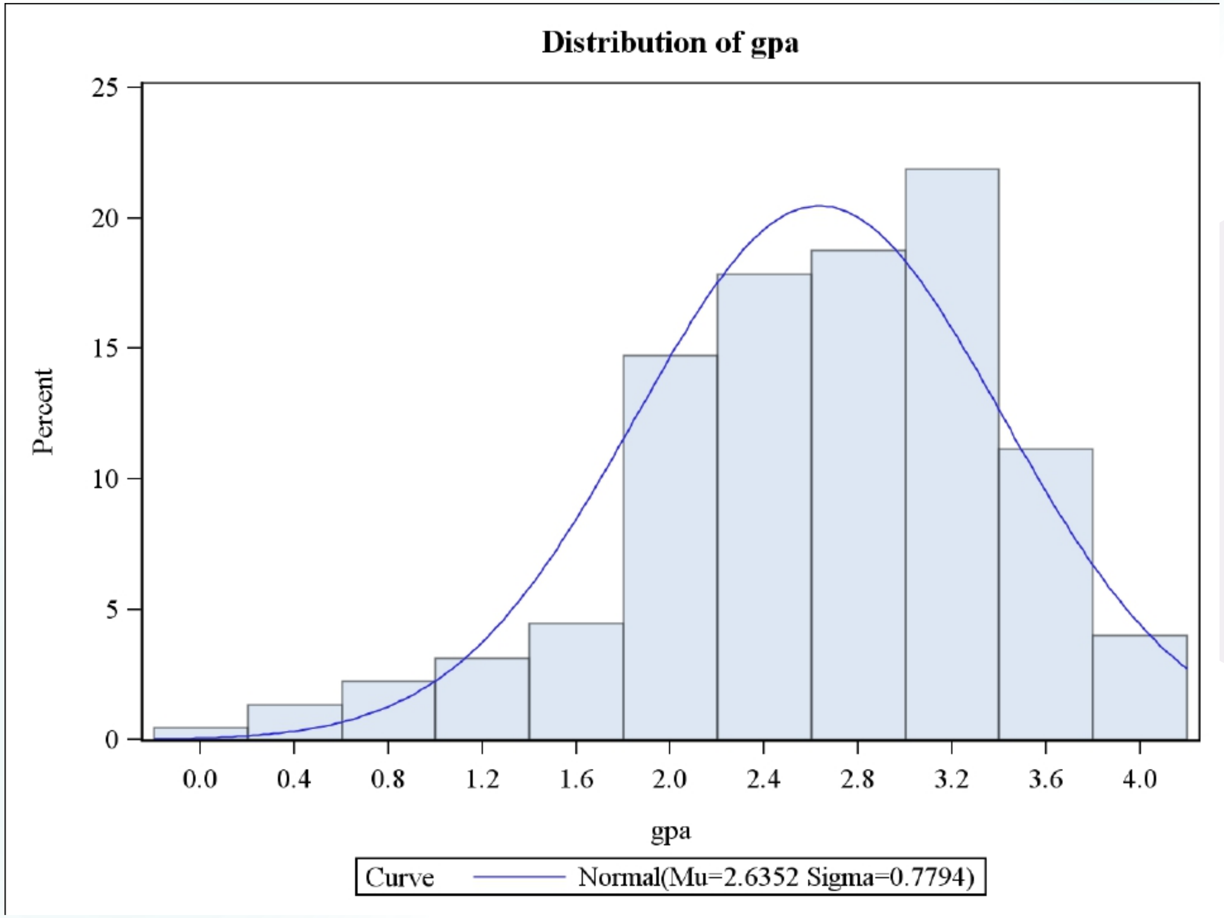
- GPA after three semesters
- Overall high school math grade
- Overall high school science grade
- Overall high school English grade
- SAT Math
- SAT Verbal
- Gender (of interest for other reasons)

- Y is the student's grade point average (GPA) after 3 semesters
- 3 HS grades and 2 SAT scores are the explanatory variables (p = 6)
- Have n = 224 students

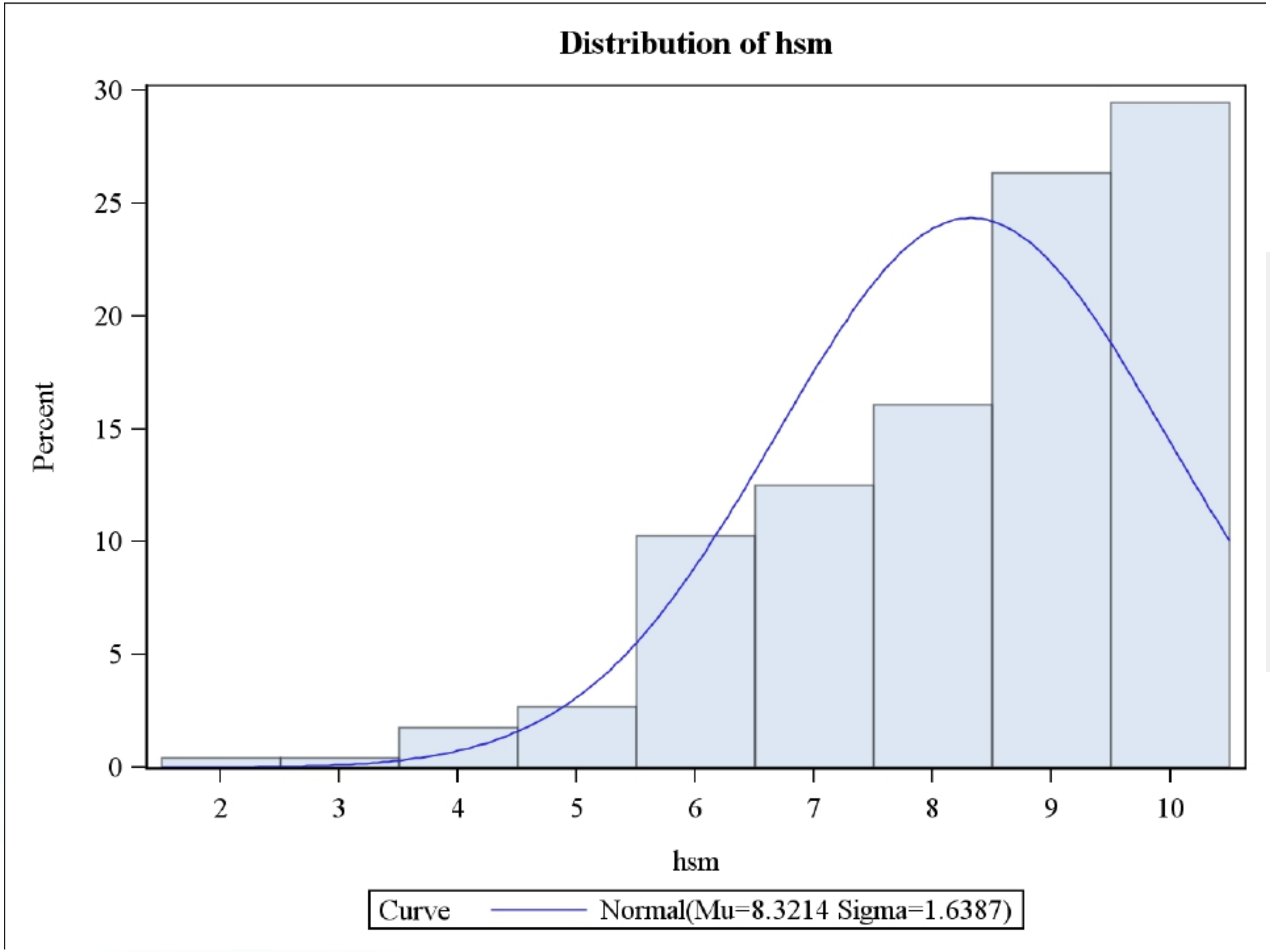
Descriptive Statistics


```
      id              gpa            hsm              hss
 Min.   :  1.00   Min.   :0.120   Min.   : 2.000   Min.   : 3.000
 1st Qu.: 56.75   1st Qu.:2.167   1st Qu.: 7.000   1st Qu.: 7.000
 Median :112.50   Median :2.740   Median : 9.000   Median : 8.000
 Mean   :112.50   Mean   :2.635   Mean   : 8.321   Mean   : 8.089
 3rd Qu.:168.25   3rd Qu.:3.212   3rd Qu.:10.000   3rd Qu.:10.000
 Max.   :224.00   Max.   :4.000   Max.   :10.000   Max.   :10.000
      hse              satm            satv             sex
 Min.   : 3.000   Min.   :300.0   Min.   :285.0   Min.   :1.000
 1st Qu.: 7.000   1st Qu.:540.0   1st Qu.:440.0   1st Qu.:1.000
 Median : 8.000   Median :600.0   Median :490.0   Median :1.000
 Mean   : 8.094   Mean   :595.3   Mean   :504.5   Mean   :1.353
 3rd Qu.: 9.000   3rd Qu.:650.0   3rd Qu.:570.0   3rd Qu.:2.000
 Max.   :10.000   Max.   :800.0   Max.   :760.0   Max.   :2.000
```

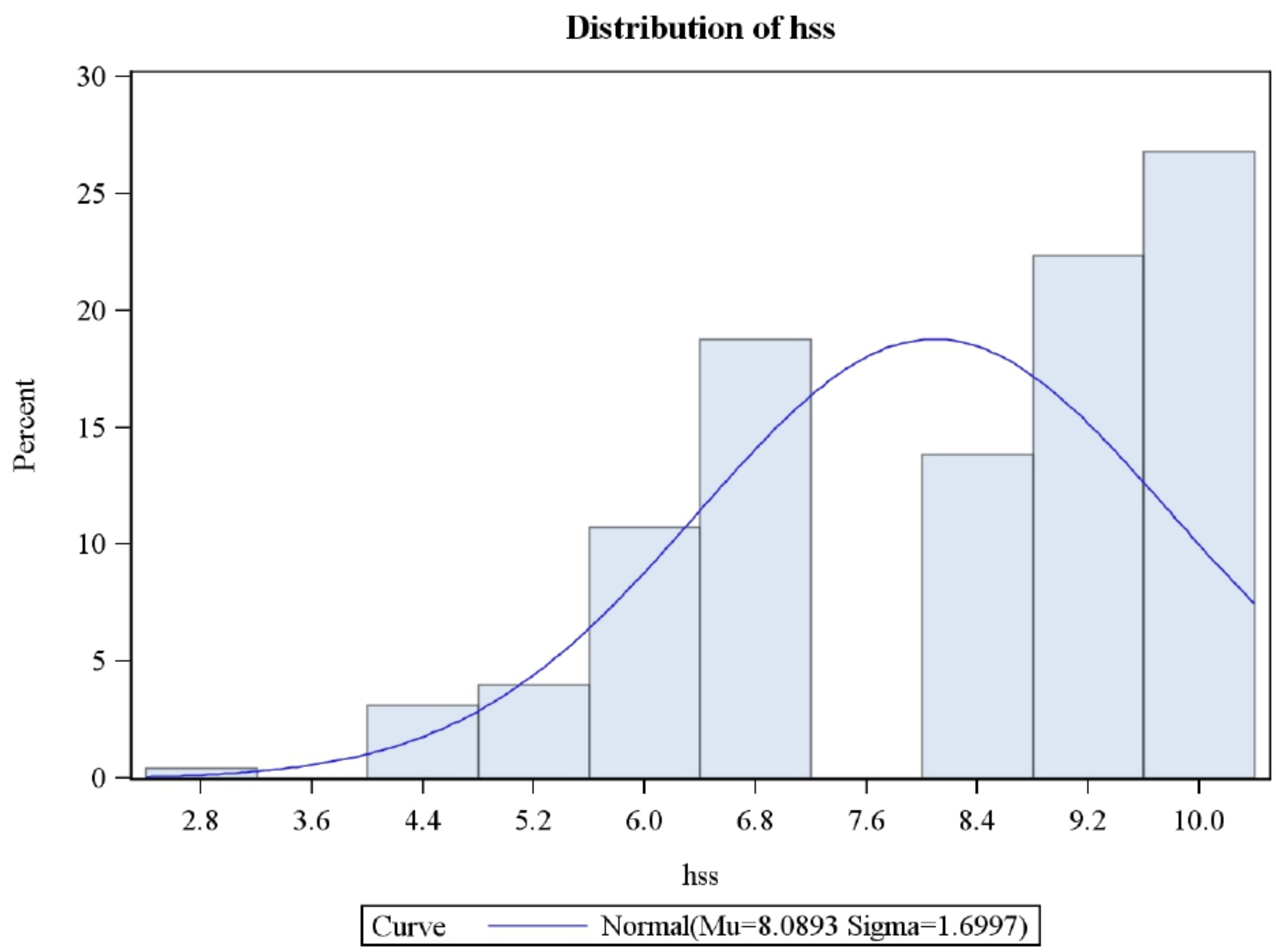
```
> summary(csdata)
```

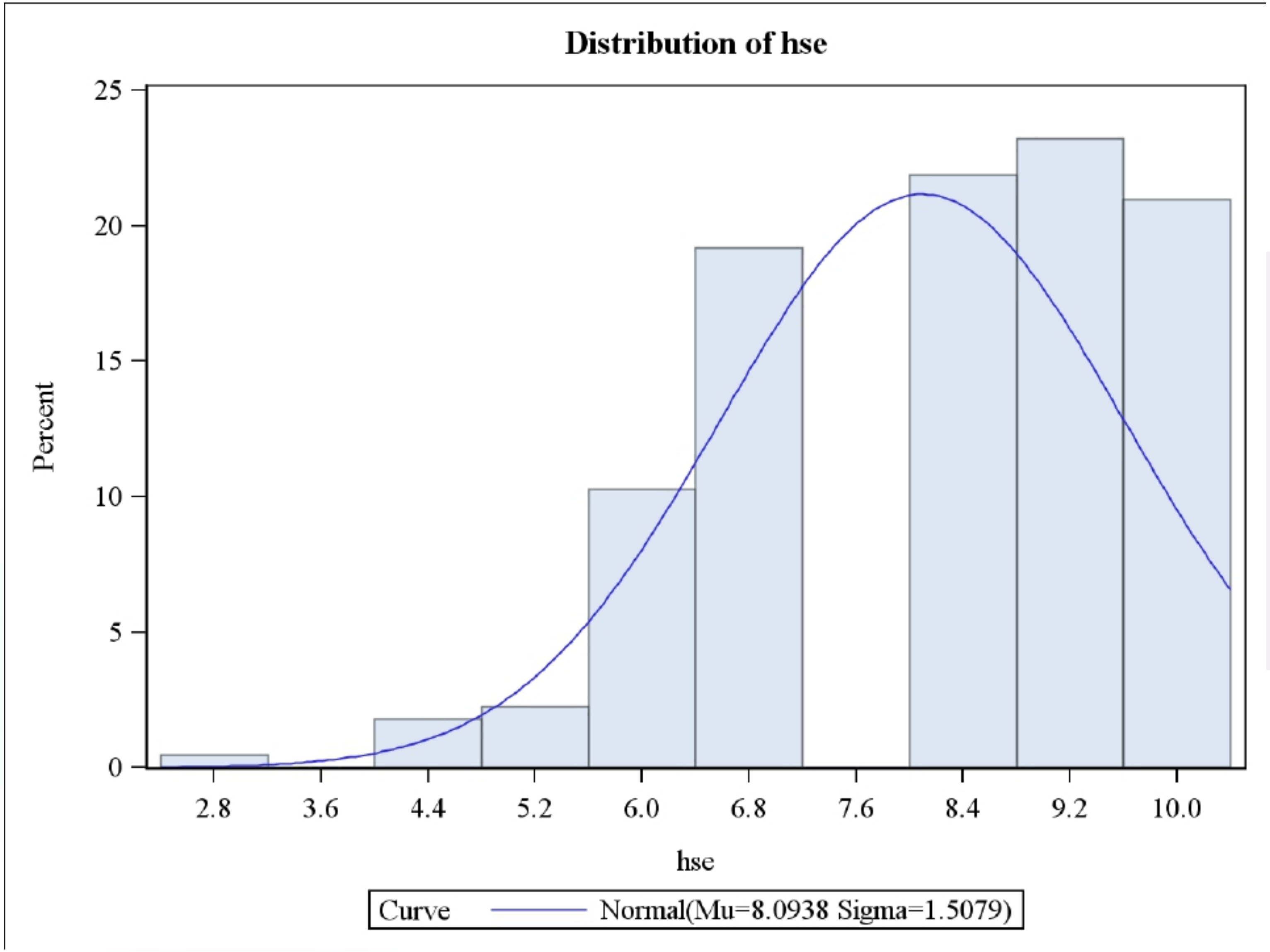


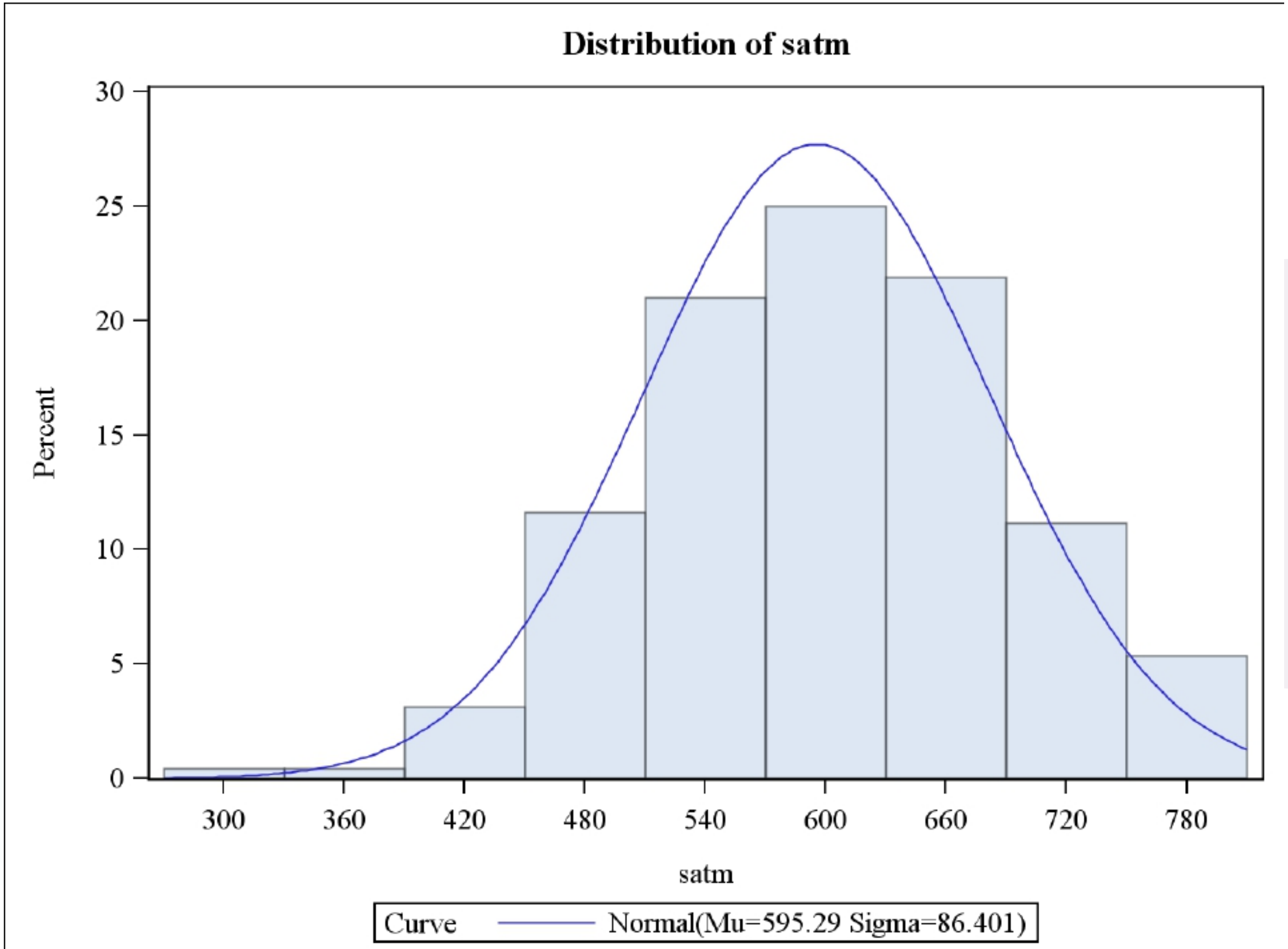
Why do we care about Y?



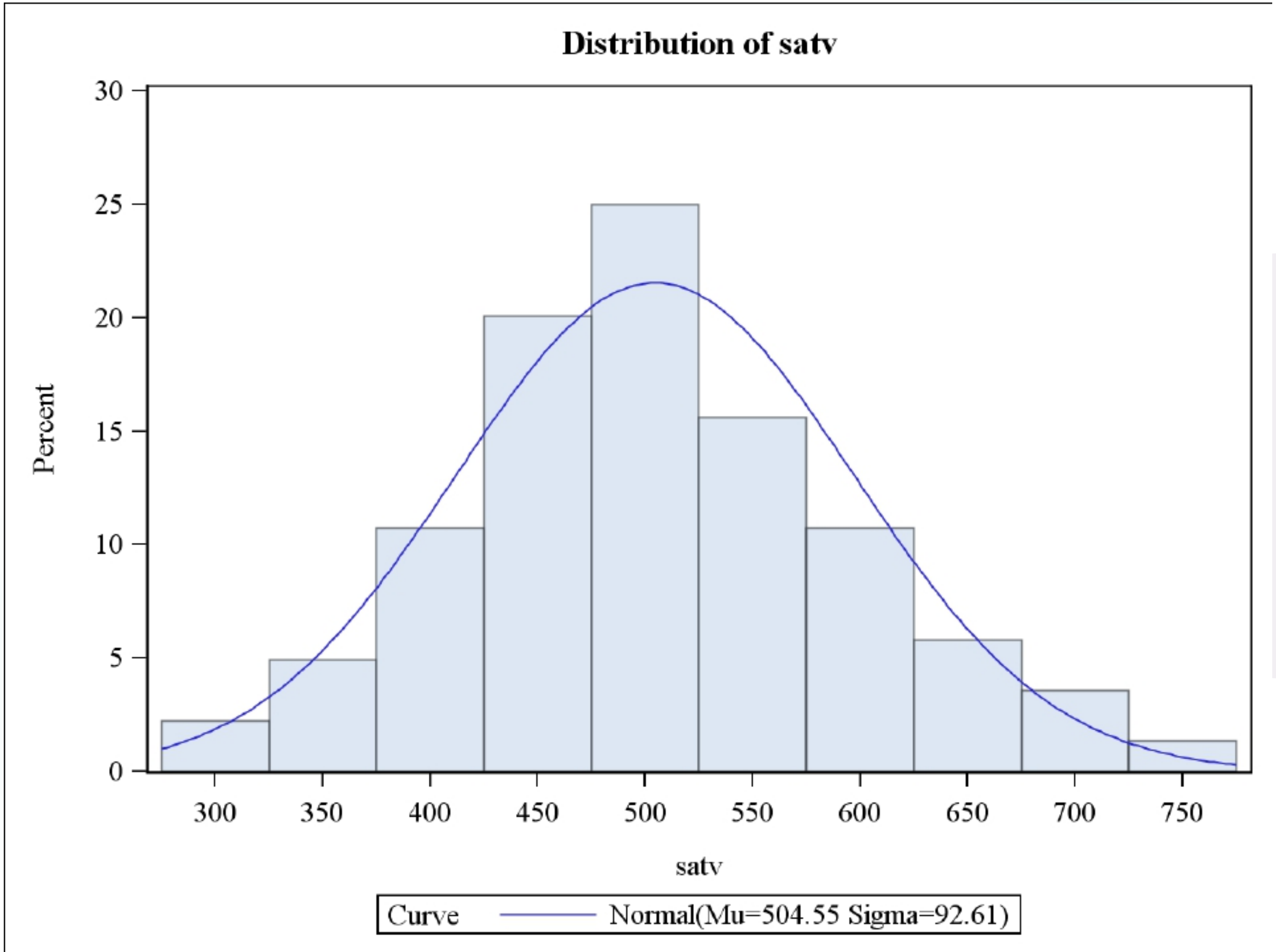
Distribution of hsm
Percent
30
25
20
15
10
5
0
2
3
4
5
6
7
8
9
10
hsm
Curve
Normal(Mu=8.3214 Sigma=1.6387)


Distribution of hss
Percent
30
25
20
15
10
5
0
2.8
3.6
4.4
5.2
6.0
6.8
7.6
8.4
9.2
10.0
hss
Curve
Normal(Mu=8.0893 Sigma=1.6997)


Distribution of hse
25
20
15
10
5
0
Percent
2.8
3.6
4.4
5.2
6.0
6.8
7.6
8.4
9.2
10.0
hse
Curve
Normal(Mu=8.0938 Sigma=1.5079)

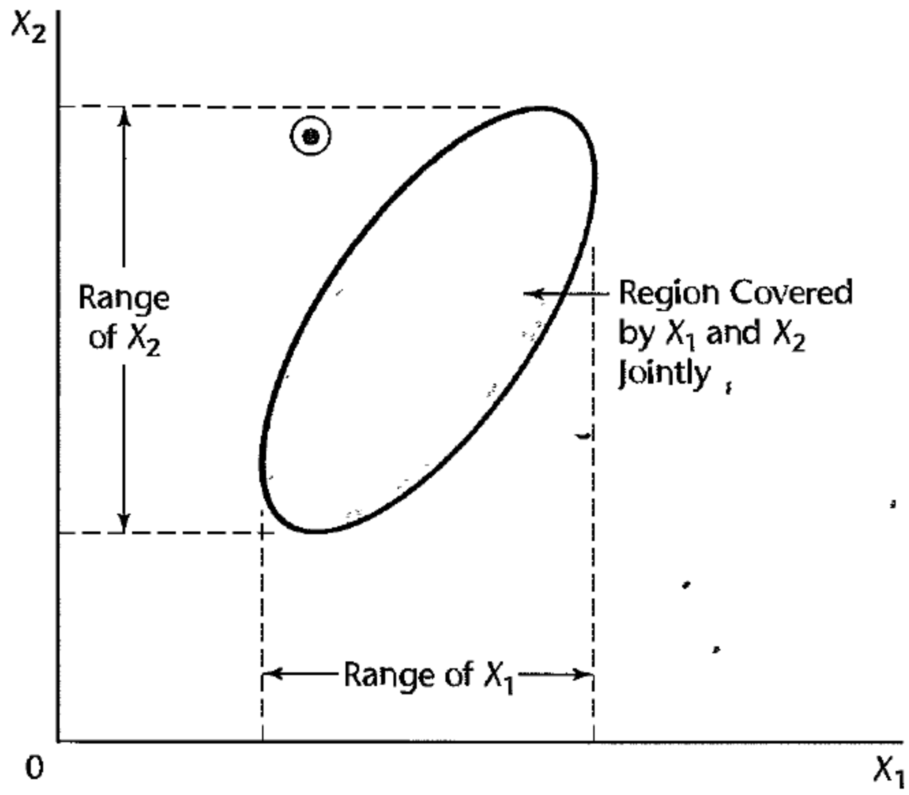


Distribution of satm
Percent
0
5
10
15
20
25
30
300
360
420
480
540
600
660
720
780
satm
Curve
Normal(Mu=595.29 Sigma=86.401)


Distribution of satv
Percent
30
25
20
15
10
5
0
300
350
400
450
500
550
600
650
700
750
satv
Curve
Normal(Mu=504.55 Sigma=92.61)

Why do We Care about X?

- Good design leads to more power in statistical inference, ensures the validity of the whole process
- Potential outliers?
- Possible confounders?
- Caution about hidden extrapolations



Correlations

Pearson Correlation Coefficients, N=224 Prob>|r|underH0:Rho=0						
	gpa	**hsm**	**hss**	**hse**	**satm**	**satv**
gpa	1.00000	0.43650 <.0001	0.32943 <.0001	0.28900 <.0001	0.25171 0.0001	0.11449 0.0873
hsm	0.43650 <.0001	1.00000	0.57569 <.0001	0.44689 <.0001	0.45351 <.0001	0.22112 0.0009
hss	0.32943 <.0001	**0.57569** **<.0001**	1.00000	0.57937 <.0001	0.24048 0.0003	0.26170 <.0001
hse	0.28900 <.0001	0.44689 <.0001	**0.57937** **<.0001**	1.00000	0.10828 0.1060	0.24371 0.0002
satm	0.25171 0.0001	0.45351 <.0001	0.24048 0.0003	0.10828 0.1060	1.00000	0.46394 <.0001
satv	0.11449 0.0873	0.22112 0.0009	0.26170 <.0001	0.24371 0.0002	0.46394 <.0001	1.00000

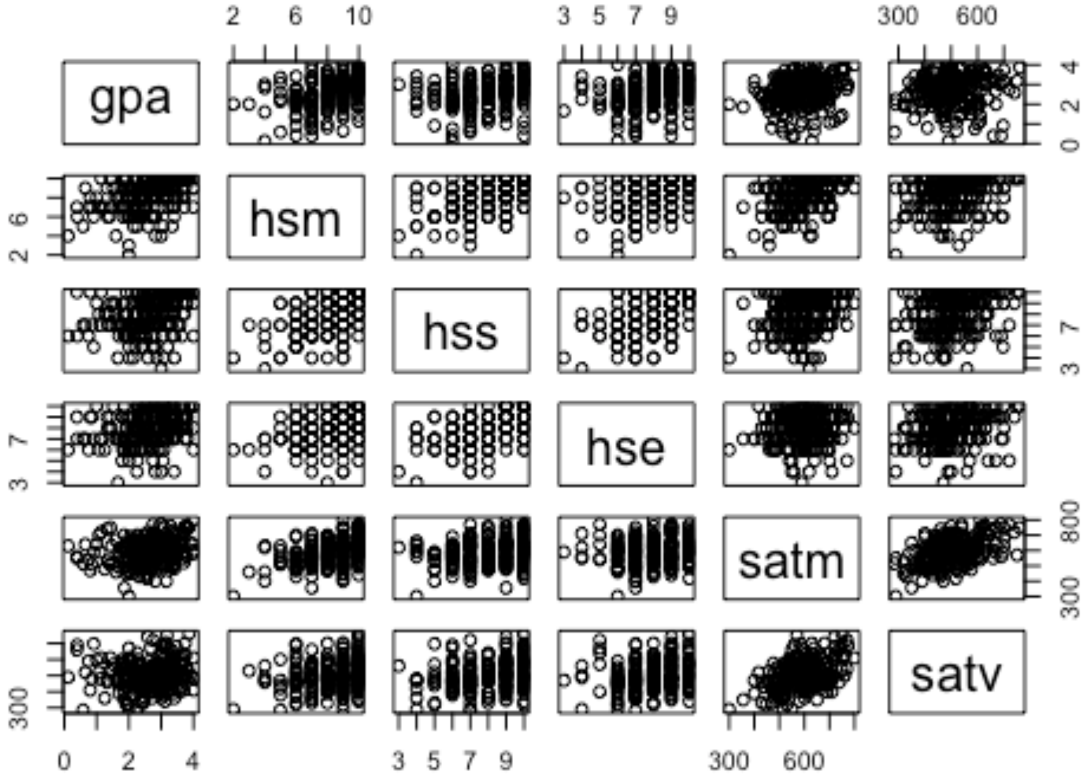
- All but satv significantly correlated with gpa

Scatter Plot Matrix

- Allows visual check of pairwise relationships
- > pairs(csdata[, c("gpa", "hsm", "hss", "hse", "satm", "satv")])

No "strong" linear Relationships

Can see discreteness of high school scores



Use High School Grades to Predict gpa (Model #1)

- fit1 = lm(gpa ~ hsm + hss + hse, data=csdata)
- summary(fit1)
- anova(fit1)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.58988	0.29424	2.005	0.0462	*
hsm	0.16857	0.03549	4.749	3.68e-06	***
hss	0.03432	0.03756	0.914	0.3619	
hse	0.04510	0.03870	1.166	0.2451	

Residual standard error: 0.6998 on 220 degrees of freedom
Multiple R-squared: 0.2046, Adjusted R-squared: 0.1937
F-statistic: 18.86 on 3 and 220 DF, p-value: 6.359e-11

Intercept Meaningful??

Analysis of Variance Table

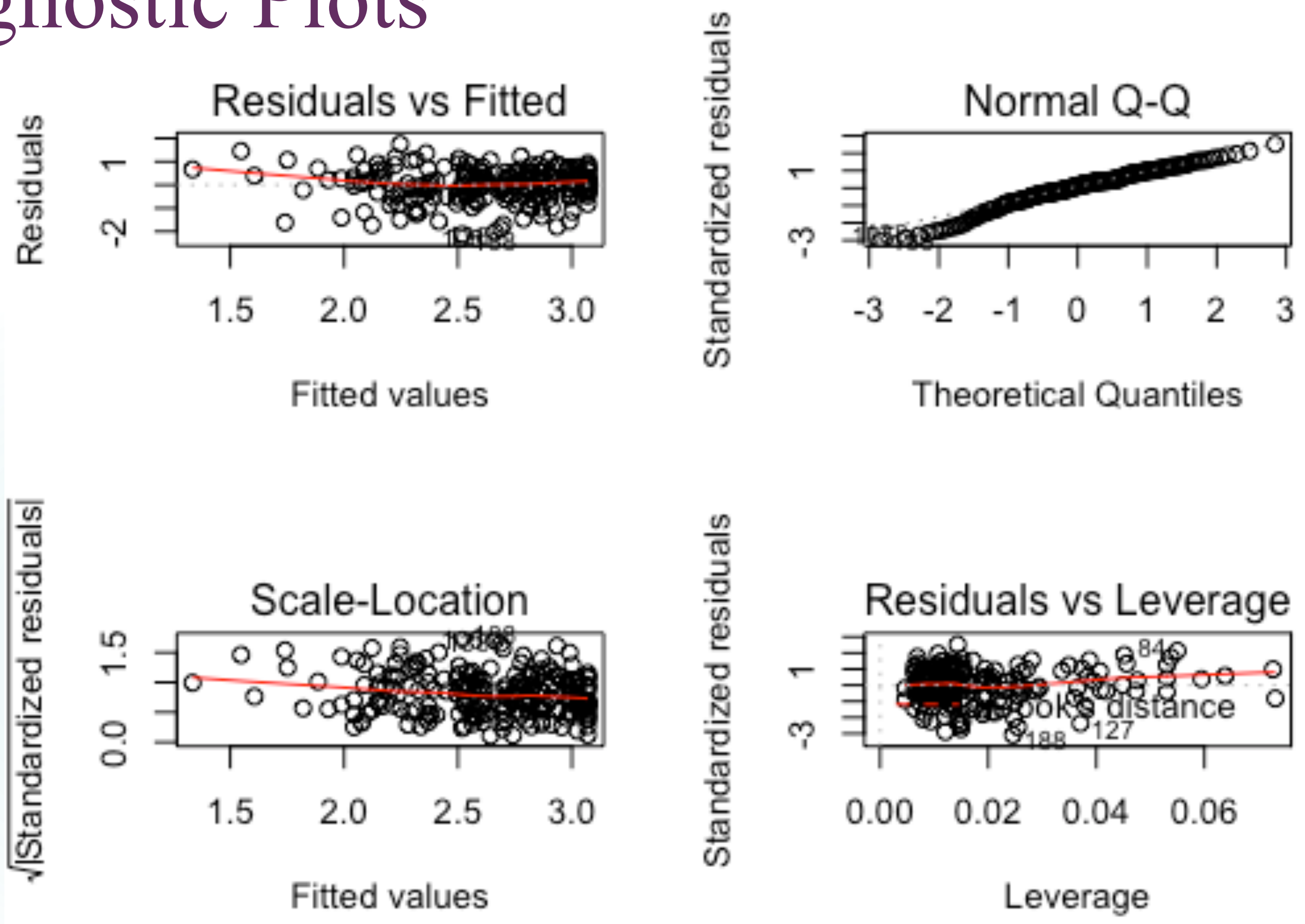
Response: gpa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
hsm	1	25.810	25.8099	52.6975	6.621e-12	***
hss	1	1.237	1.2371	2.5258	0.1134	
hse	1	0.665	0.6654	1.3585	0.2451	
Residuals	220	107.750	0.4898			

Significant F test but not all variable t tests significant

Fit Diagnostic Plots

- plot(fit1)



Remove hss (Model #2)

> fit2 = lm(gpa ~ hsm + hse, data=csdata)
> summary(fit2)
> anova(fit2)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.62423	0.29172	2.140	0.0335	*
hsm	0.18265	0.03196	5.716	3.51e-08	***
hse	0.06067	0.03473	1.747	0.0820	.

Residual standard error: 0.6996 on 221 degrees of freedom
Multiple R-squared: 0.2016, Adjusted R-squared: 0.1943
F-statistic: 27.89 on 2 and 221 DF, p-value: 1.577e-11

Model #1's Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.58988	0.29424	2.005	0.0462	*
hsm	0.16857	0.03549	4.749	3.68e-06	***
hss	0.03432	0.03756	0.914	0.3619	
hse	0.04510	0.03870	1.166	0.2451	

Residual standard error: 0.6998 on 220 degrees of freedom
Multiple R-squared: 0.2046, Adjusted R-squared: 0.1937
F-statistic: 18.86 on 3 and 220 DF, p-value: 6.359e-11

Analysis of Variance Table

Response: gpa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
hsm	1	25.810	25.8099	52.7369	6.443e-12	***
hse	1	1.494	1.4936	3.0518	0.08203	.
Residuals	221	108.159	0.4894			

Slightly better MSE and adjusted R-Sq

Significant F test but not all variable t tests significant

Rerun with hsm Only (Model #3)

- fit3 = lm(gpa ~ hsm, data=csdata)
- summary(fit3)
- anova(fit3)
- plot(fit3)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.90768    0.24355   3.727 0.000246 ***
hsm          0.20760    0.02872   7.229 7.77e-12 ***

Residual standard error: 0.7028 on 222 degrees of freedom
Multiple R-squared:  0.1905,  Adjusted R-squared:  0.1869
F-statistic: 52.25 on 1 and 222 DF,  p-value: 7.774e-12
```

Slightly worse MSE and adjusted R-Sq

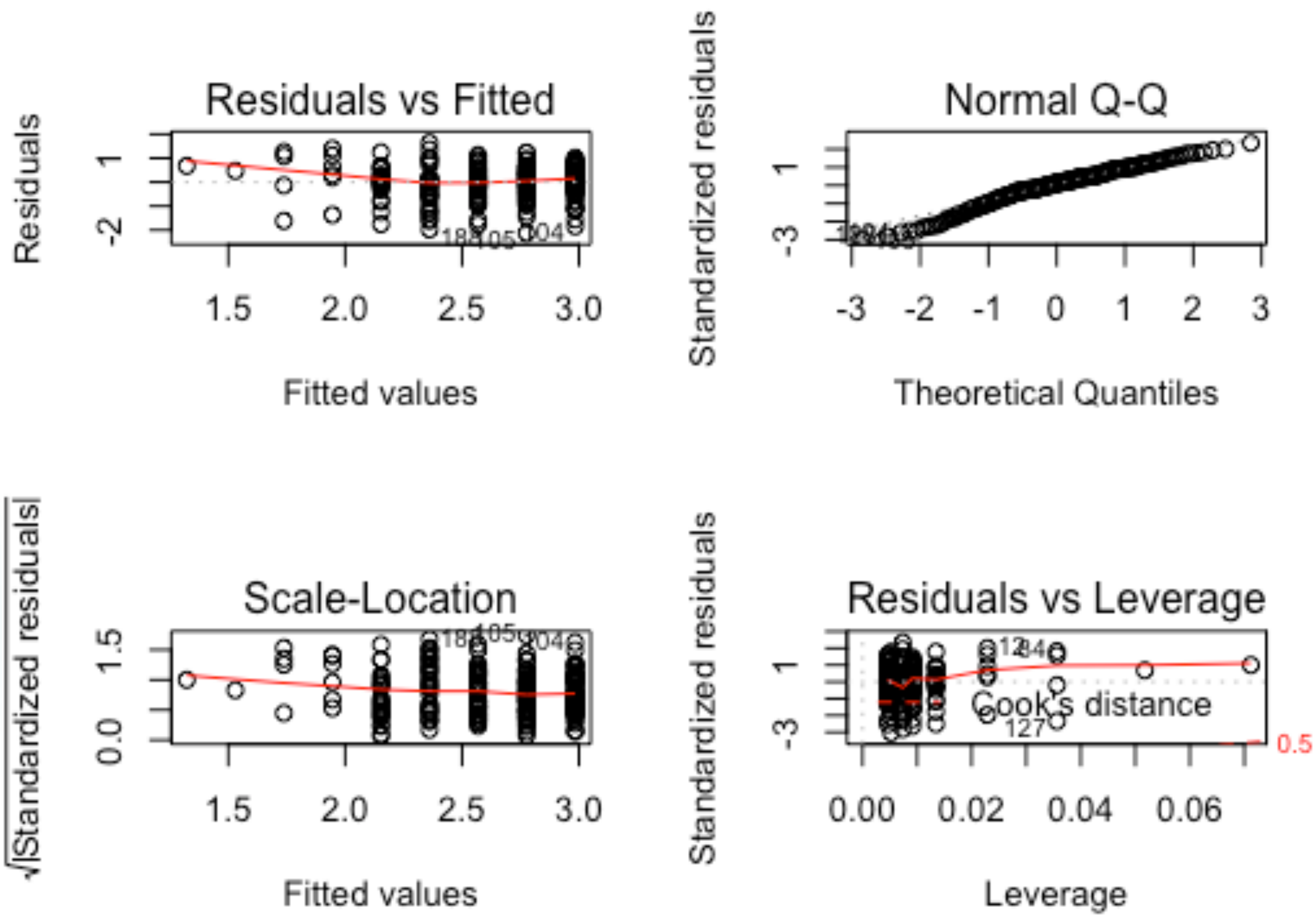
```
Model #2's Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.62423    0.29172   2.140   0.0335 *
hsm          0.18265    0.03196   5.716 3.51e-08 ***
hse          0.06067    0.03473   1.747   0.0820 .

Residual standard error: 0.6996 on 221 degrees of freedom
Multiple R-squared:  0.2016,  Adjusted R-squared:  0.1943
F-statistic: 27.89 on 2 and 221 DF,  p-value: 1.577e-11
```

```
Analysis of Variance Table

Response: gpa
           Df Sum Sq Mean Sq F value    Pr(>F)
hsm         1  25.81 25.8099  52.254 7.774e-12 ***
Residuals 222 109.65  0.4939
```

Significant F test and all variable t tests significant



Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
Leverage

SATs (Model #4)

> fit4 = lm(gpa ~ satm + satv, data=csdata)

> summary(fit4)

> anova(fit4)

> plot(fit4)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.289e+00	3.760e-01	3.427	0.000728	***
satm	2.283e-03	6.629e-04	3.444	0.000687	***
satv	-2.456e-05	6.185e-04	-0.040	0.968357	

Residual standard error: 0.7577 on 221 degrees of freedom
Multiple R-squared: 0.06337, Adjusted R-squared: **0.05489**
F-statistic: 7.476 on 2 and 221 DF, p-value: 0.0007218

Analysis of Variance Table

Response: gpa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
satm	1	8.583	8.5829	14.9499	0.0001452	***
satv	1	0.001	0.0009	0.0016	0.9683570	
Residuals	221	126.879	0.5741			

Significant F test but not all variable t tests significant

Much worse MSE and adjusted R-Sq

HS and SATs (Model #5)

- fit5 = lm(gpa ~ hsm + hss + hse + satm + satv, data=csdata)
- summary(fit5)
- anova(fit5)
- plot(fit5)

```
Coefficients:
              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  0.3267187  0.3999964   0.817   0.414932
hsm          0.1459611  0.0392610   3.718   0.000256 ***
hss          0.0359053  0.0377984   0.950   0.343207
hse          0.0552926  0.0395687   1.397   0.163719
satm         0.0009436  0.0006857   1.376   0.170176
satv        -0.0004078  0.0005919  -0.689   0.491518

Residual standard error: 0.7 on 218 degrees of freedom
Multiple R-squared:  0.2115,	Adjusted R-squared: 
0.1934
F-statistic: 11.69 on 5 and 218 DF,  p-value: 5.058e-10
```

Model Comparisons

Cannot reject the reduced model…
No significant information lost…
We don't need SAT variables

\# test for satm and satv

- reduced1 = lm(gpa ~ hsm + hss + hse, data=csdata)
- anova(reduced1, fit5)

```
Analysis of Variance Table

Model 1: gpa ~ hsm + hss + hse
Model 2: gpa ~ hsm + hss + hse + satm + satv
  Res.Df  RSS    Df  Sum of Sq     F     Pr(>F)
1     220  107.75
2     218  106.82  2     0.93131   0.9503 0.3882
```

\# test for hsm + hss + hse

- reduced2 = lm(gpa ~ satm + satv, data=csdata)
- anova(reduced2, fit5)

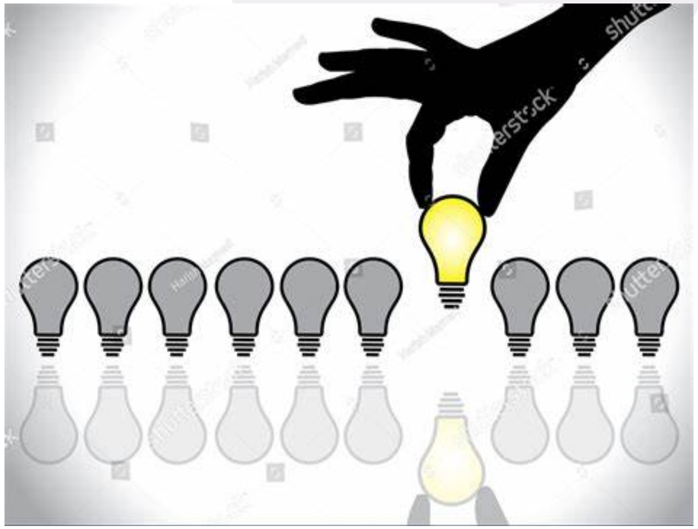
```
Analysis of Variance Table

Model 1: gpa ~ satm + satv
Model 2: gpa ~ hsm + hss + hse + satm + satv
  Res.Df   RSS Df Sum of Sq     F      Pr(>F)
1   221   126.88
2   218   106.82  3     20.06    13.646 3.432e-08 ***
```

Reject the reduced model…There is significant information lost… We can't remove HS variables from model

Best Model?

- Likely the one with just HSM or the one with HSE and HSM (Model #2)
- We'll discuss model selection and comparison methods in Chapters 7 and 8



Key Ideas from Case Study

- First, look at graphical and numerical summaries one variable at a time
- Then, look at relationships between pairs of variables with graphical and numerical summaries
- Use plots and correlations to understand relationships
- The relationship between a response variable and an explanatory variable depends on what other explanatory variables are in the model
- A variable can be a significant (P-value<.05) predictor alone and not significant (P-value>.05) when other X's are in the model
- Regression coefficients, standard errors and the results of significance tests depend on what other explanatory variables are in the model

Key Ideas from Case Study

- Significance tests (P values) do not tell the whole story
- Squared multiple correlations (R^2, give the proportion of variation in the response variable explained by the explanatory variables) can give a different view
- We often express R^2 as a percent
- You can fully understand the theory in terms of $Y = X\beta + \varepsilon$
- However to effectively use this methodology in practice you need to understand how the data were collected, the nature of the variables, and how they relate to each other

Background Reading

- lec8_cs2.R contains the R commands used in this topic

Topic 2:

Inference in Multiple Regression

Outline

- Review Multiple Linear Regression
- Inference of Regression Coefficients
  - Application to book example
- Inference of Mean
  - Application to book example
- Inference of Future Observation
- Diagnostics and Remedies

Multiple Regression

- Data
  - Y is the response variable
  - $X_1, X_2, \ldots, X_{p\text{-}1}$ are the p-1 explanatory variables
  - $Y_i, X_{i1}, X_{i2}, \ldots, X_{i,p\text{-}1}$ are the data for case i where $i = 1$ to n
- Model

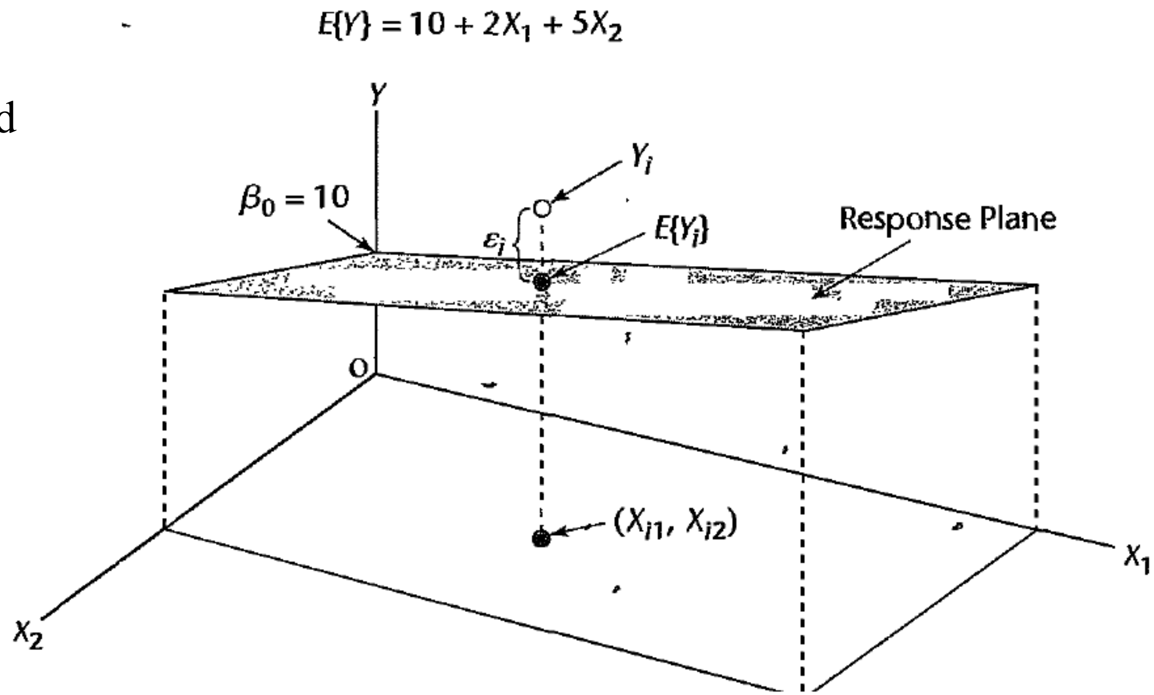
$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

  - Y_i is the value of the response variable for the case
  - β_0 is the intercept
  - $\beta_1, \beta_2, \ldots, \beta_{p\text{-}1}$ are the regression coefficients for the explanatory variables
  - ε_i's are independent Normally distributed random errors with mean 0 and variance σ^2

Geometric Illustration

- The regression function is called a regression surface or a response surface
- The parameter β_1 indicates the change in the mean response E{Y} per unit increase in X_1 when X_2 is held constant
- The first-order regression model is designed for predictor variables whose effects on the mean response are additive or do not interact
- Then the response function is a plane
- β_1 and β_2 are called partial regression coefficients because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$



Least Squares Solutions

$$b = (X'X)^{-1}X'Y$$

$$\hat{Y} = Xb = X(X'X)^{-1}X'Y = HY$$

$$e = Y - \hat{Y} = (I - H)Y$$

$$s^2 = \frac{e'e}{n-p} = \frac{Y'(I-H)Y}{n-p}$$

$$s = \text{ root MSE } = \sqrt{s^2}$$

- b and e are independent, therefore, b and s^2 are independent under normal error terms

ANOVA F Test for Linear Relation

- Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

$$H_1 : \beta_k \neq k \text{ for at least one } k \text{ in } 1, 2, \ldots, p-1.$$

- Test statistic:

$$F^* = \frac{MSM}{MSE}$$

- Sampling distribution under H_0:

$$F^* \sim F_{p-1,n-p}$$

- Decision rule at α

1. Reject H_0 if the calculated $F_0 > F_{p\text{-}1,n\text{-}p,\alpha}$
2. Reject H_0 if the P-value $P(F^* > F_0 | H_0) < \alpha$

Inference for Individual Coefficients

- Recall the sampling distribution of b:

$$b = \left(b_0, b_1, b_2, \ldots, b_{p-1}\right)' \sim N\left(\beta,\ \sigma^2 (X'X)^{-1}\right)$$

- Define

$$s^2(b)_{p\times p} = s^2(X'X)^{-1} = MSE(X'X)^{-1}$$

- For b_k:

$$s^2(b_k) = s^2(b)_{k,k} = MSE\left((X'X)^{-1}\right)_{k,k}$$

the k^{th} diagonal entry

Significance Test for β_k

- Hypotheses:

$$H_0: \beta_k = 0 \quad vs \quad H_1: \beta_k \neq 0$$

- Test Statistic:

$$t^* = \frac{b_k}{s(b_k)}$$

- Sampling distribution under

$$t^* \sim t_{df_E} = t_{n-p}$$

- Decision rules: the *P*-value and the critical value approaches as before
- This tests the significance of explanatory variable X_k given the other variables in the model

Confidence Interval for β_k

- From:

$$b_k \sim N(\beta_k, \sigma^2((X'X)^{-1})_{k,k})$$

$$\frac{(n-p)s^2}{\sigma^2} = \frac{e'e}{\sigma^2} \sim \chi^2_{n-p}$$

b_k and $s(b_k)$ (Standard Error of b_k) are independent

- We have, under the model:

$$\frac{b_k - \beta_k}{s(b_k)} \sim t_{n-p}$$

- 100(1-α)% Confidence Interval for β_k

$$b_k \pm t_{\alpha/2,n-p} s(b_k)$$

Note: Proof of $\frac{e'e}{\sigma^2} \sim \chi^2_{n-p}$

- <u>Theorem</u>: $X \sim N(\mu, I_p)$, A is symmetric, then

  $$X'AX \sim \chi^2_{r,\ \mu'A\mu} \Leftrightarrow A \text{ is idempotent and rank}(A)=r$$

- Proof of $\frac{e'e}{\sigma^2} \sim \chi^2_{n-p}$:
  - $Y \sim N(X\beta, \sigma^2 I_n)$, $\therefore e = Y - \hat{Y} = (I-H)Y \sim N\big(0, \sigma^2(I-H)\big)$,
  - $e^* = \frac{1}{\sigma}(I-H)^{-\frac{1}{2}} e \sim N(0, I)$, $e = \left[\sigma(I-H)^{\frac{1}{2}}\right] e^*$,
  - $\therefore e'e = e^{*\prime}[\sigma^2(I-H)]e^* \sim \chi^2_r$

    where r = rank(I-H) = tr(I-H) = n-p

Studio Example (KNNL p 236)

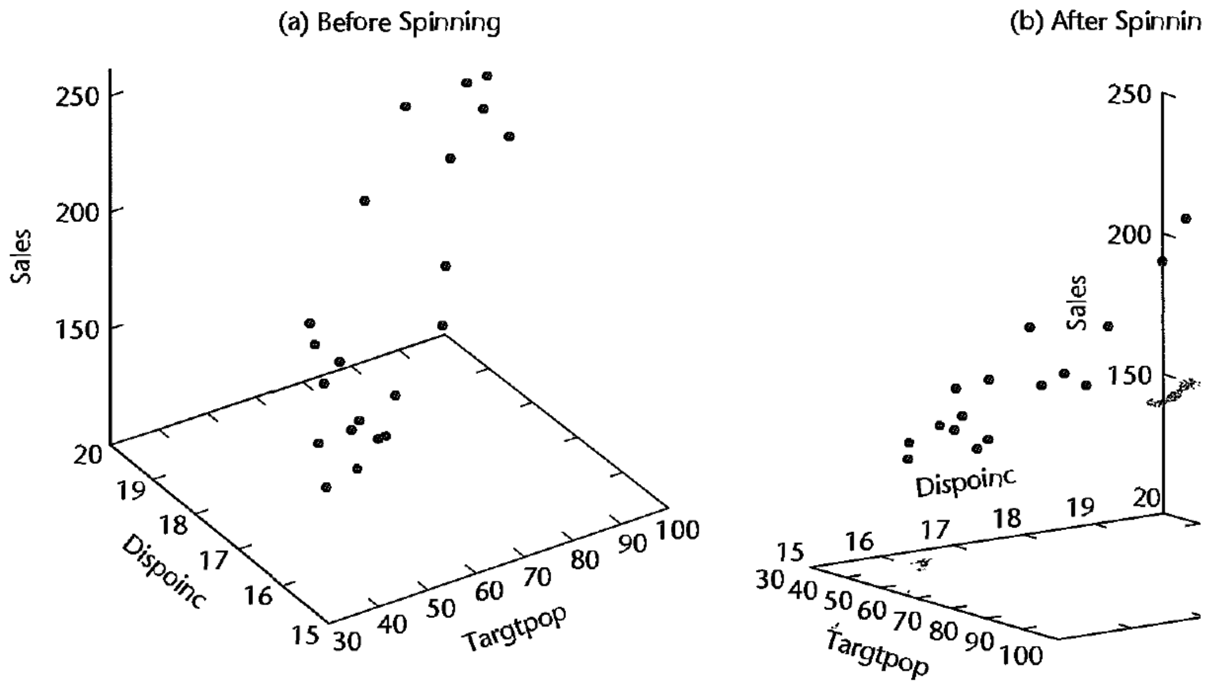
- Dwaine Studios, Inc. operates portrait studios in 21 cities of medium size
- The company is considering an expansion into other cities of medium size and wishes to investigate whether sales in a community can be predicted from the number of persons aged 16 or younger in the community and the per capita disposable personal income in the community
  - Y: the total sale in a city
  - X_1: population aged 16 and under (thousands)
  - X_2: per capita disposable income (thousands)
- The model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Read in the Data

```
> a <- file.choose() #choose "CH06FI05.txt"
> a1 <- read.table(a)
> colnames(a1) <- c("Targtpop", "Dispoinc", "Sales")
> head(a1)
```

	Targtpop	Dispoinc	Sales
1	68.5	16.7	174.4
2	45.2	16.8	164.4
3	91.3	18.2	244.2
4	47.8	16.3	154.6
5	46.9	17.3	181.6
6	66.1	18.2	207.56

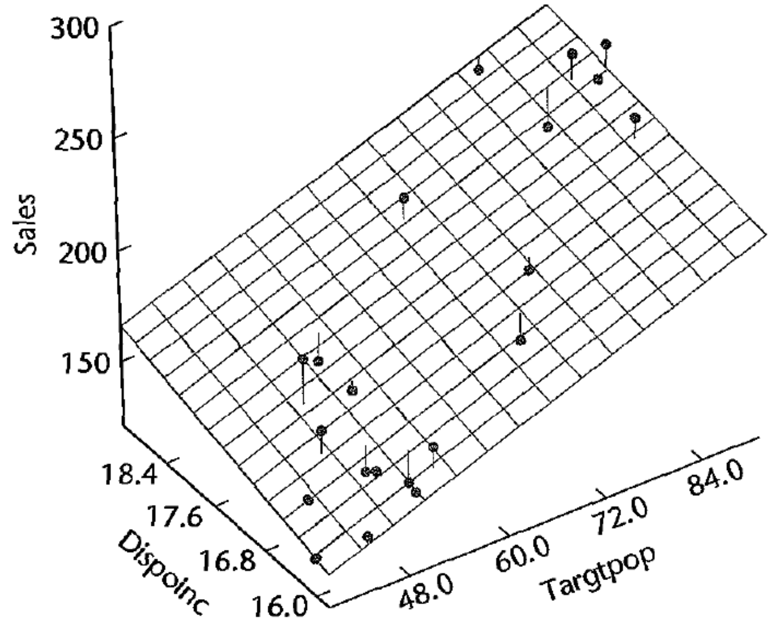


Regression

- reg1 <- lm(Sales ~ Targtpop + Dispoinc, data=a1)
- summary(reg1)
- anova(reg1)

Both variables are helpful in explaining Sales when the other is already in the model

ParameterEstimates				
	Parameter Estimate	Standard Error	t Value	Pr>|t|
(Intercept)	-68.85707	60.01695	-1.15	0.2663
Targtpop	1.45456	0.21178	6.87	<.0001
Dispoinc	9.36550	4.06396	2.30	0.0333



ANOVA Output

Analysis of Variance Table

Response: Sales

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Targtpop	1	23371.8	23371.8	192.8962	4.64e-11	***
Dispoinc	1	643.5	643.5	5.3108	0.03332	*
Residuals	18	2180.9	121.2			

Root MSE	11.00739	**R-Square**	0.917

At least one variable is helpful in predicting in Sales

Confidence Intervals

- Use confint( ) to get confidence intervals for each coefficient

```
> confint(reg1, level=0.95)
```

Output:

```
                   2.5 %     97.5 %
(Intercept) -194.9480130 57.233867
Targtpop       1.0096226  1.899497
Dispoinc       0.8274411 17.903560
```

What if Just Include Targtpop?

ParameterEstimates					
Variable	DF	Parameter Estimate	Standard Error	95%Confidence Limits	
Intercept	1	68.04536	9.46224	48.24066	87.85006
Targtpop	1	1.83588	0.14641	1.52943	2.14233

- CIs for both the intercept and Targtpop change dramatically when just Targtpop as explanatory variable
- Coefficients depend on other variables in model

	2.5 %	97.5 %
(Intercept)	-194.9480130	57.233867
Targtpop	1.0096226	1.899497
Dispoinc	0.8274411	17.903560

Estimation of Mean Response $E(Y_h)$

- $X_{h.}$ is now a vector that looks like

$$(1, X_{h1}, X_{h2}, \ldots, X_{h,p\text{-}1})$$

- We want a point estimate and a confidence interval for the mean response $E(Y_h)$ corresponding to the set of explanatory variables $X_{h.}$

Inference Theory for $E(Y_h)$

- $\mu_h = E(Y_h) = X_{h.}\beta$
- Estimator:

$$\hat{\mu}_h = X_{h.}b = X_{h.}(X'X)^{-1}X'Y$$

- Sampling distribution of $\hat{\mu}_h$:

$$\hat{\mu}_h \sim N(\mu_h, \sigma^2 X_{h.}(X'X)^{-1}X'_{h.})$$

- Estimated variance:

$$s^2(\hat{\mu}_h) = s^2 X_{h.}(X'X)^{-1}X'_{h.}$$

- 100(1- α)% Confidence Interval for μ_h:

$$\hat{\mu}_h \pm t_{\frac{\alpha}{2},n-p}\, s\,(\hat{\mu}_h)$$

Using predict()

> conf_interval = predict(reg1, se.fit = TRUE, interval="confidence", level = 0.95)

OutputStatistics							
Obs	Targtpop	Dispoinc	Dependent Variable	Predicted Value	StdError se.fit	95%CLMean lwr	upr
1	68.5	16.7	174.4000	187.1841	3.8409	179.114	195.2536
2	45.2	16.8	164.4000	154.2294	3.5558	146.759	161.6998
3	91.3	18.2	244.2000	234.3963	4.5882	224.756	244.0358
4	47.8	16.3	154.6000	153.3285	3.2331	146.536	160.1210
5	46.9	17.3	181.6000	161.3849	4.4300	152.077	170.6921
21	52.3	16.0	166.5000	157.0644	4.0792	148.494	165.6344

Prediction of New Y_h

- $X_{h.}$ is still a vector of form

$$(1, X_{h1}, X_{h2}, \ldots, X_{h,p\text{-}1})$$

- We want a prediction of Y_h based on a set of predictor values with an interval that expresses all of the uncertainty in our prediction
- Uncertainty = Uncertainty from sample + New error term

Inference Theory for Y_h

- $Y_h = X_{h.}\beta + \varepsilon$
- Predictor:

$$\hat{Y}_h = X_{h.}b = X_{h.}(X'X)^{-1}X'Y$$

- Distribution of $\hat{Y}_h - Y_h$:

$$\hat{Y}_h - Y_h \sim N(0, \sigma^2 + \sigma^2 X_{h.}(X'X)^{-1}X_{h.}')$$

- Estimated variance:

$$s^2(\hat{Y}_h - Y_h) = s^2[1 + X_{h.}(X'X)^{-1}X_{h.}']$$

- 100(1- α)% Confidence Interval for Y_h:

$$\hat{Y}_h \pm t_{\frac{\alpha}{2}, n-p}\, s(\hat{Y}_h - Y_h)$$

Note on $\hat{Y}_h - Y_h \sim N(0, \sigma^2 + \sigma^2 X_{h.}(X'X)^{-1}X_{h.}')$

$$Y_h - \hat{Y}_h = (X_h\beta + \varepsilon_h) - X_h\hat{\beta}$$

$$= (X_h\beta + \varepsilon_h) - X_h\big((X'X)^{-1}X'(X\beta + \varepsilon)\big)$$

$$= \varepsilon_h - X_h(X'X)^{-1}X'\varepsilon$$

$$\because\ \varepsilon_h\ and\ \varepsilon\ are\ independent, \therefore\ \varepsilon_h - X_h(X'X)^{-1}X'\varepsilon \sim N(0, \upsilon)$$

$$where\quad \upsilon = Var(\varepsilon_h) + Var(X_h(X'X)^{-1}X'\varepsilon\,)$$

$$= \sigma^2(1 + X_h(X'X)^{-1}X_h{}')$$

Using predict()

```
> conf_interval = predict(reg1, se.fit = TRUE, interval="predict", level = 0.95)
```

OutputStatistics							
Obs	**Targtpop**	**Dispoinc**	**Dependent Variable**	**Predicted Value**	**StdError se.fit**	**95%CLMean lwr**	**upr**
1	68.5	16.7	174.4000	187.1841	3.8409	162.691	211.6772
2	45.2	16.8	164.4000	154.2294	3.5558	129.927	178.531
3	91.3	18.2	244.2000	234.3963	4.5882	209.342	259.450
21	52.3	16.0	166.5000	157.0644	4.0792	132.401	181.727

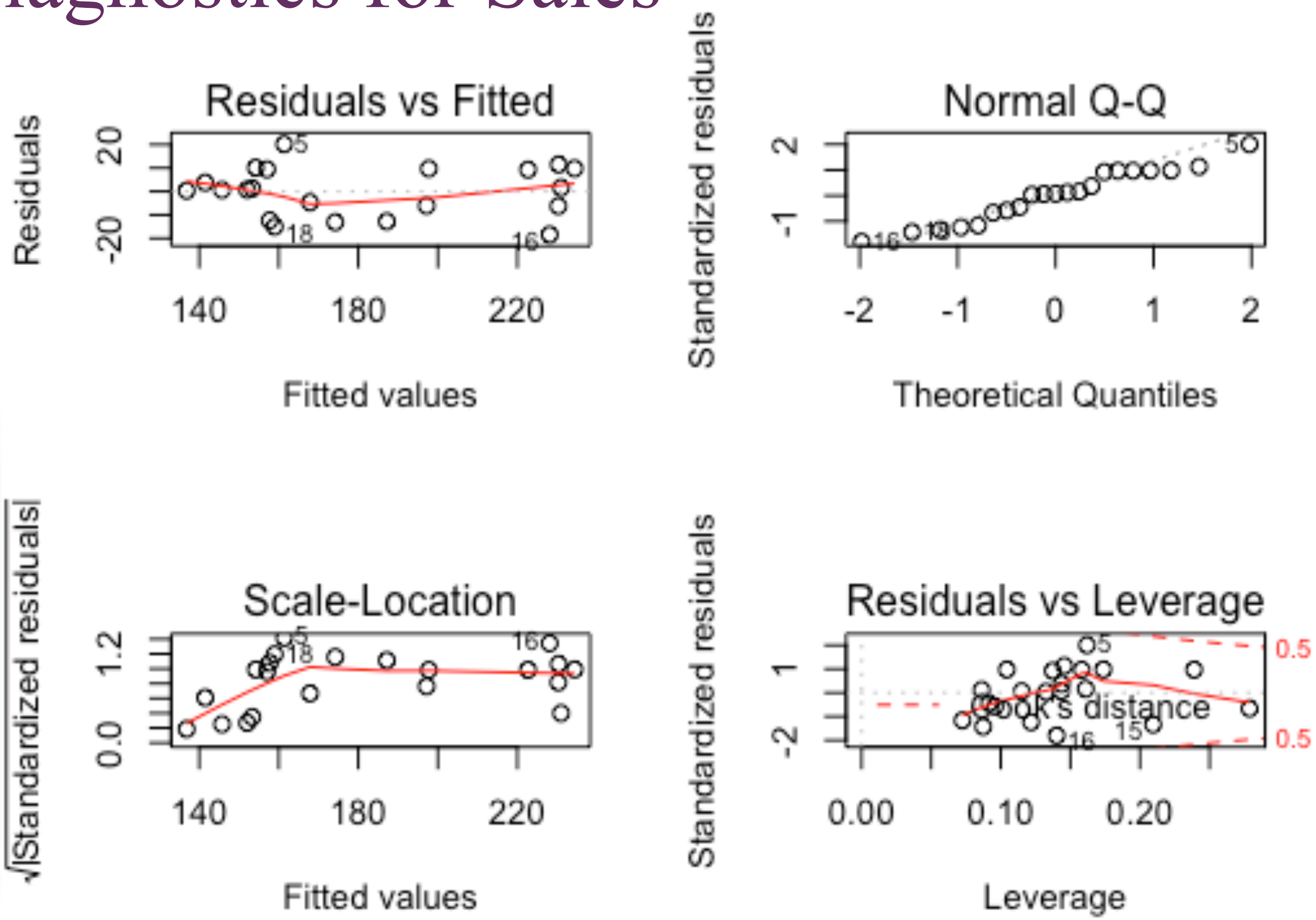
Diagnostics

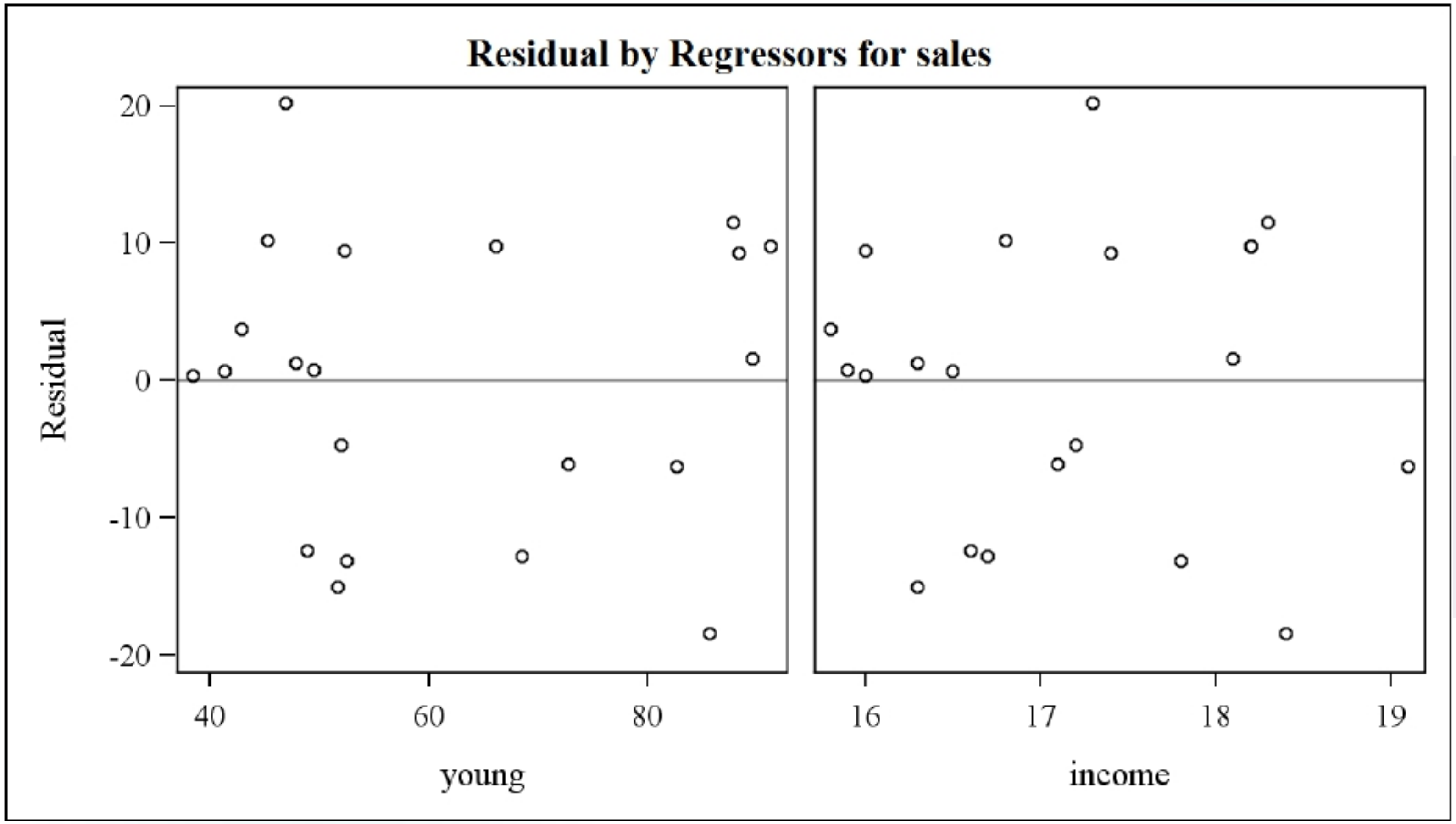
- Look at the distribution of each variable
- Look at the relationship between pairs of variables
- Plot the residuals versus
  - the predicted/fitted values
  - each explanatory variable
  - time or order (if available)

Diagnostics

- Are the residuals approximately Normal?
  - Look at the histogram
  - Normal quantile plot
- Is the variance constant?
- Plot the residuals vs anything that might be related to the variance (e.g. residuals *vs* predicted values & residuals versus each *X*)

Fit Diagnostics for Sales




Residual by Regressors for sales
Residual
20
10
0
-10
-20
40
60
80
young
16
17
18
19
income

Remedies

- Similar remedies as simple regression
- Transformations such as Box-Cox
- Analyze with/without possible outliers
- More details to come in Chapters 9 and 10

Background Reading

- We finished Chapter 6
- Program used to generate output for confidence intervals for means and prediction intervals is lec8.R