

$$\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_n \end{bmatrix}$$

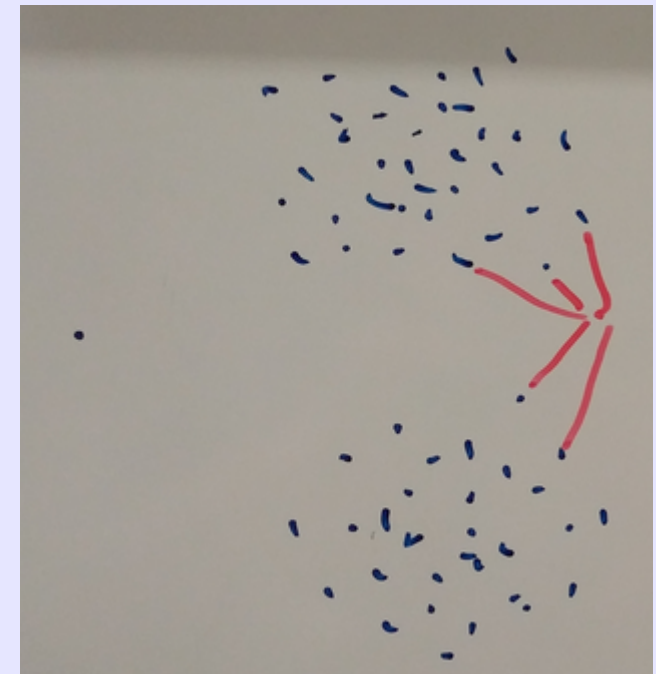
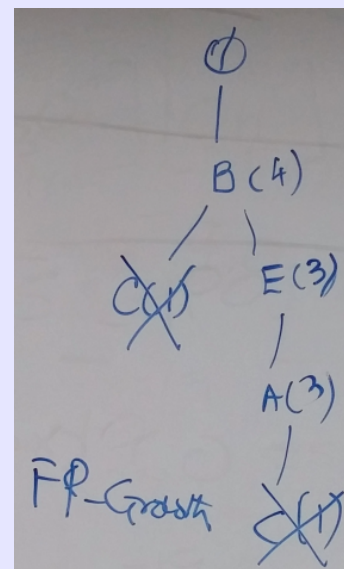
$$X = \sum_{i=1}^{\text{rank}(X)} \sigma_i u_i v_i^T = U \Sigma V^T$$

σ_i : i^{th} singular value of X
 u_i : i^{th} left singular value of X (i^{th} column of U)
 v_i^T : i^{th} right singular vector of X (i^{th} column of V^T)

Captures the patterns among attributes
 Captures the patterns among the objects

CS 422: Data Mining
 Vijay K. Gurbani, Ph.D.,
 Illinois Institute of Technology

Lecture: **Linear regression**



Linear regression: Example

- Let's check the effect of radio advertising on the sales through linear regression:

```
> model.radio <- lm(sales ~ radio, data=df)
> summary(model.radio)

Call:
lm(formula = sales ~ radio, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-15.7305  -2.1324   0.7707   2.7775   8.1810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542  <2e-16 ***
radio        0.20250    0.02041   9.921  <2e-16 ***

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

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,    Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

Interpretation:

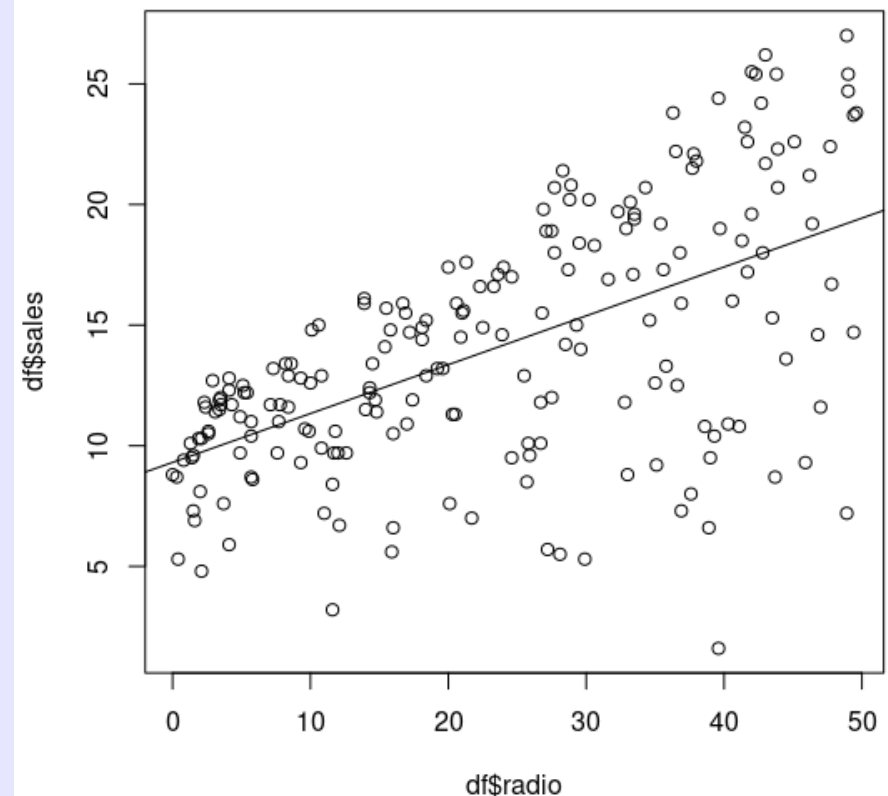
A \$1000 increase in spending on radio advertisement yields an average increase in sales of about 212 units.

- Regression equation: $\text{sales} = \beta_0 + \beta_1 \cdot \text{radio}$
 $= 9.312 + 0.203 \cdot \text{radio}$

Linear regression: Example

- Let's see how the regression line looks like for a single regressor (radio):

```
> plot(df$radio, df$sales)  
> abline(model.radio)
```



Linear regression: Example

- Let's check the effect of all advertising media on the sales:

```
> model <- lm(sales ~ ., data=df)
> summary(model)
```

Call:
lm(formula = sales ~ ., data = df)

Residuals:

	Min	1Q	Median	3Q	Max
	-8.8277	-0.8908	0.2418	1.1893	2.8292

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.938889	0.311908	9.422	<2e-16 ***
TV	0.045765	0.001395	32.809	<2e-16 ***
radio	0.188530	0.008611	21.893	<2e-16 ***
newspaper	-0.001037	0.005871	-0.177	0.86

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

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

- Regression equation:

$$\begin{aligned}\text{sales} &= \beta_0 + \beta_1 * \text{TV} + \beta_2 * \text{radio} + \beta_3 * \text{newspaper} \\ &= 2.939 + 0.046 * \text{TV} + 0.189 * \text{radio} - 0.001 * \text{newspaper}\end{aligned}$$

Linear regression: Hypothesis testing

- p-values, statistical significance (α), and hypothesis tests.
 - H_0 (Null hypothesis) vs. H_1 (or H_a , alternative hypothesis)
 - α : $P(\text{rejecting } H_0 \mid H_0 \text{ is true})$
 - p-value: $P(\text{getting a result as extreme as you have} \mid H_0 \text{ is true})$
- Hypothesis test:
 - p-value $\leq \alpha$: result does not support H_0 ,
result statistically significant.
 - p-value $> \alpha$: result supports H_0 .

Linear regression: Hypothesis testing

- p-values, statistical significance (α), and hypothesis tests.
 - H_0 (Null hypothesis) vs. H_1 (or H_a , alternative hypothesis)
 - α : $P(\text{rejecting } H_0 \mid H_0 \text{ is true})$
 - p-value: $P(\text{getting a result as extreme as you have} \mid H_0 \text{ is true})$
- Hypothesis test:
 - p-value $\leq \alpha$: result does not support H_0 , **result statistically significant.**
 - p-value $> \alpha$: result supports H_0 .

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "reject H_0 in favor of H_1 " or "do not reject H_0 "; we never conclude "reject H_1 ", or even "accept H_1 ".

Linear regression: Analysis

- To understand the *fit* of a regression model, we need to ask some important questions.
 - 1) Is at least one of the predictors useful in predicting the response?
 - 2) Do all the predictors help explain the response, or is only a subset of predictors useful?
 - 3) How well does the model fit the data?
 - 4) Given a set of predictor values (the β 's), what response value should we predict and how accurate is our prediction?

Linear regression: Analysis

1. Is at least one of the predictors useful in predicting the response?

$$H_0 = \beta_1 = \beta_2 = \dots \beta_p = 0$$

$$H_1 = \text{At least one } \beta \text{ is non zero}$$

To answer the above question regarding H_0 and H_1 , we define a F-statistic:

If there is no relationship between response and predictors, F-statistic is close to 1 (H_0 is true).

If H_1 is true, F-statistic > 1 .

The *p-value* associated with the F-statistic is very small, giving us strong evidence that one of the predictors is associated with increased sales.

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871  -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

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

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

$$TSS: \sum_{i=1}^n (y_i - \bar{y})^2$$

Linear regression: Analysis

2. Do all the predictors help explain the response, or is only a subset of predictors useful?

As before, we observe the p-values of the predictors in context of the following hypothesis test:

$H_0 : \beta_i = 0$ (There is no relationship between predictor and response)

$H_1 : \beta_i \neq 0$ (There is some relationship between predictor and response)

The p-values for TV and radio appear to imply that there is a relationship between these predictor and sales (p-values low).

The p-value for newspaper appears to imply that there may not be a relationship between newspaper and sales; i.e., H_0 cannot be rejected.

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871  -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Linear regression: Analysis

Let's drill down into the newspaper.

Residuals:

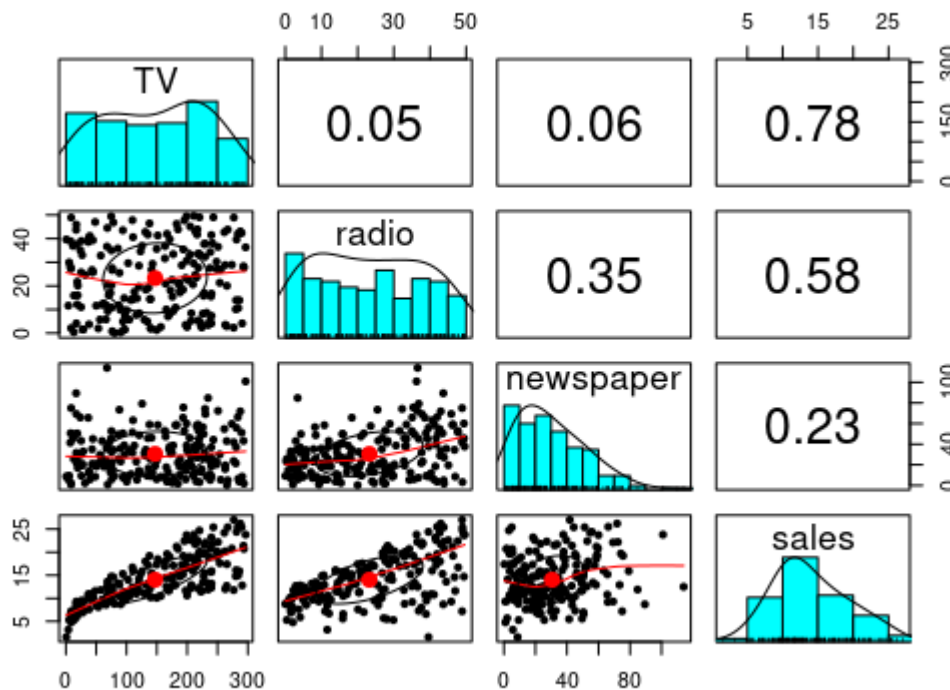
Min	1Q	Median	3Q	Max
-8.8277	-0.8908	0.2418	1.1893	2.8292

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.938889	0.311908	9.422	<2e-16 ***
TV	0.045765	0.001395	32.809	<2e-16 ***
radio	0.188530	0.008611	21.893	<2e-16 ***
newspaper	-0.001037	0.005871	-0.177	0.86

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

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16



Linear regression: Analysis

Generally, what we are trying to do is *feature selection*, which is a hard problem.

Given p predictors, we will have 2^p models! (> 1 million models for $p=30$!)

Strategies:

- Forward selection: Start with null model (intercept) and add predictors.
- Backward selection: Start with all predictors and remove variables with largest p-values.

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871  -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Linear regression: Analysis

3. How well does the model fit the data?

The following statistics describe this: Residual Standard Error (RSE), R^2 , and RMSE.

$$RSE = \sqrt{\frac{1}{n - p - 1} RSS}$$

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}}$$

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871  -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

$$R^2 = \text{Cor}(Y, \hat{Y})^2$$

R^2 for model that uses all three predictors = 0.8972

R^2 for model that uses TV and radio only = 0.8971

Linear regression: Analysis

3. How well does the model fit the data?

The following statistics describe this:

Residual Standard Error (RSE), R^2 and RMSE.

$$RSE = \sqrt{\frac{1}{n - p - 1} RSS}$$

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}}$$

$$R^2 = Cor(Y, \hat{Y})^2$$

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871   -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
> anova(model)
Analysis of Variance Table

Response: sales
      Df Sum Sq Mean Sq    F value    Pr(>F)
TV      1 3314.6   3314.6 1166.7308 <2e-16 ***
radio    1 1545.6   1545.6  544.0501 <2e-16 ***
newspaper 1    0.1     0.1    0.0312  0.8599
Residuals 196  556.8     2.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear regression: Analysis

3. How well does the model fit the data?

The following statistics describe this:
Residual Standard Error (RSE), R^2 and RMSE.

$$RSE = \sqrt{\frac{1}{n - p - 1} RSS}$$

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}}$$

$$R^2 = Cor(Y, \hat{Y})^2$$

RSE	Predictors
3.26	TV
1.681	TV + radio
1.686	TV + radio + newspaper

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871  -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

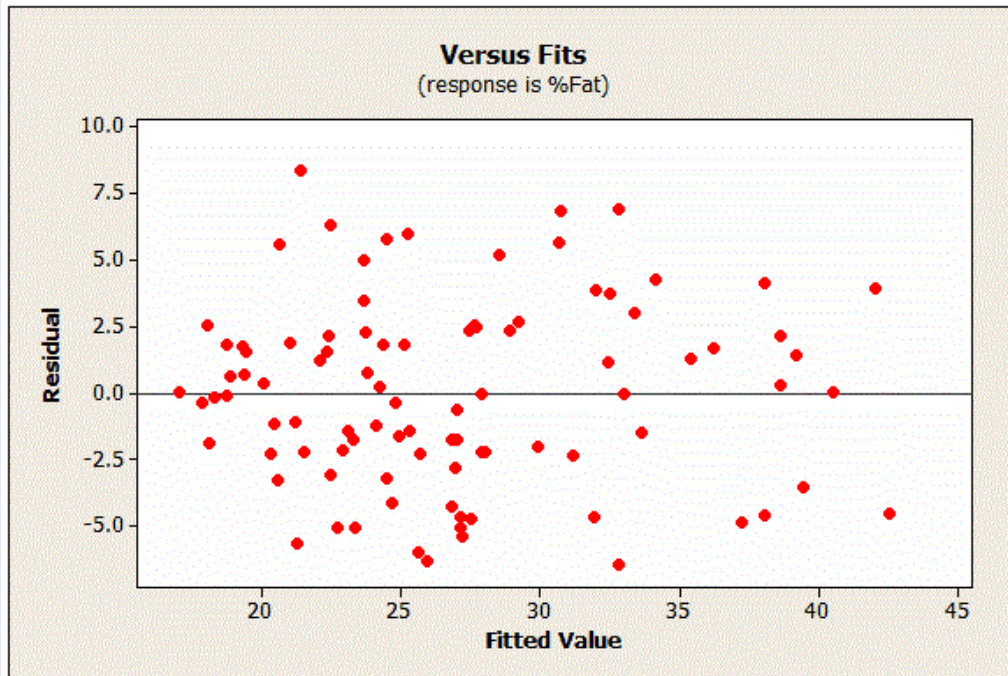
Linear regression: Analysis

Residual analysis is one more tool to see how well does the model fit the data.

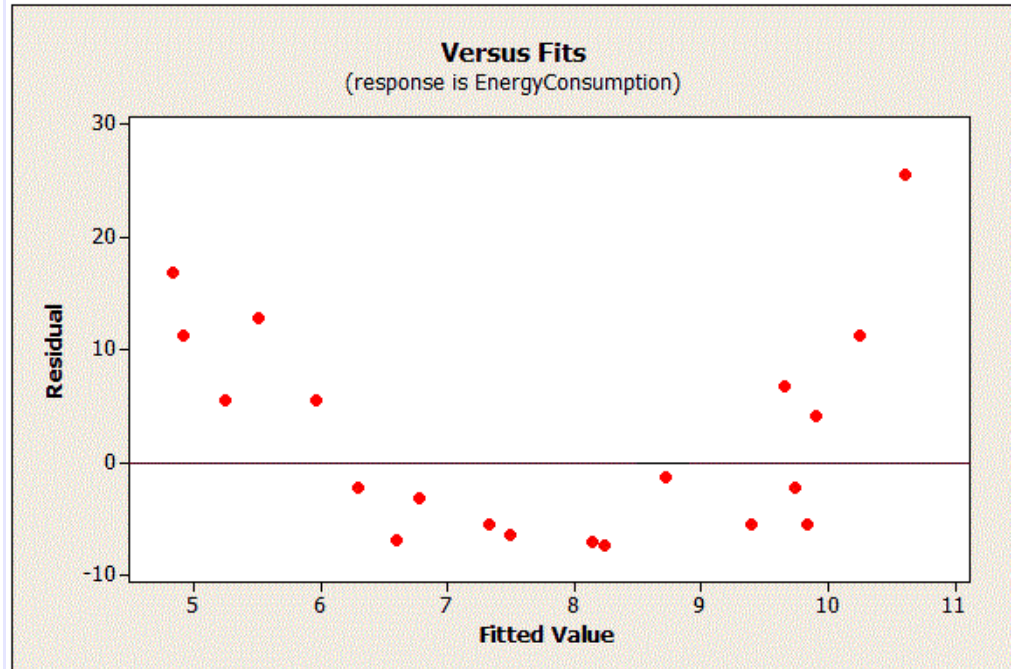
- Residuals should be homoscedastic (variance in the residual must not vary too much between low and high values of the fitted values).
- Centered on 0 through the range of fitted values ($\mu = 0.0$).
- Residuals should be normally distributed.
- Residuals must be uncorrelated to each other.

Linear regression: Analysis

Graphic source: <http://blog.minitab.com/blog/adventures-in-statistics-2/why-you-need-to-check-your-residual-plots-for-regression-analysis>



A good residual plot



A problematic residual plot

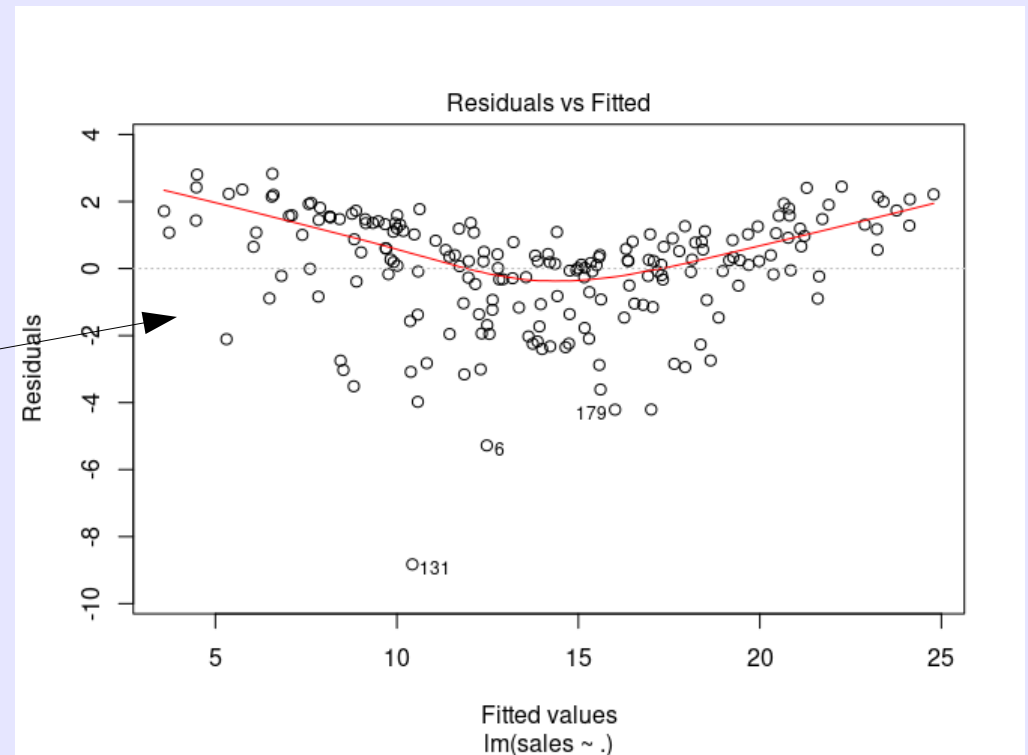
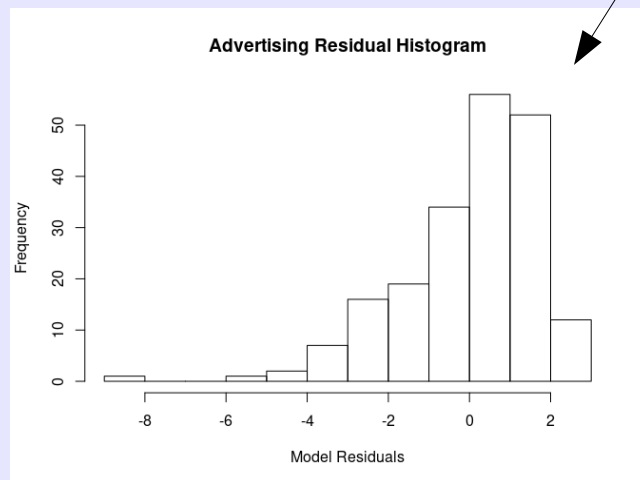
We can predict the residuals for fitted values. Residuals for 6.5-9.9 are negative, while those for 5, 6, and > 10 are positive.

Linear regression: Analysis

What do the residual plot from the Advertising data look like?

```
# Plotting residuals ...
plot(model, 1) # The easy way using plot()
# Or the manual way
plot(model$fitted.values, model$residuals,
     xlab = "Fitted values\nlm(sales ~.)",
     ylab="Residuals",
     main="Residuals vs. Fitted");
abline(0, 0)
```

- Looks reasonably good as residuals appear homoscedastic and clustered around 0. (Though they don't appear to be normally distributed.)



- Note a slight convex shape (concave upward), which may indicate some non-linearity in the data.

Linear regression: Analysis

4. Given a set of predictor coefficients (the β 's), what response value should we predict and how accurate is our prediction?

- The least squares plane of a fitted model is an *approximation* to the true population regression plane: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_n X_n \approx Y = f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$
- The approximation introduces some error.
- To account for the error, we compute confidence intervals in order to determine how close \hat{Y} will be to $f(X)$.

Linear regression: Analysis

4. Given a set of predictor values (the β 's), what response value should we predict and how accurate is our prediction?

To understand this, we need to understand:

- Significance level (α): Probability of making the wrong decision, given H_0 is true.
- Confidence interval: The range of results that would be expected to contain the population parameter of interest.
- Confidence level: Probability that if an experiment was repeated multiple times, results will be the same.
Confidence level = $1 - \alpha$.

Linear regression: Analysis

Confidence Interval

Gallup poll before 2020 elections: Two in three Americans (66%) saying prior to the election "...that they are "very" or "somewhat confident" that votes will be cast and counted accurately across the country." The margin of sampling error was ± 6 percentage points at the 95% confidence level.

(Here ± 6 percentage points is the Margin of Error.)

Linear regression: Analysis

Confidence Level

Probability that if an experiment was repeated multiple times, results will be the same.

Confidence level = $1 - \alpha$.

95% Confidence level means if the poll was to be repeated 20 times, Gallup would expect similar results 19 times ($19/20 = 0.95$).

What is the relationship between Confidence Level and Confidence Interval?

Linear regression: Analysis

```
Console ~/IIT/CS422/lectures/regression/ ↗  
> # Let's see how we predict using the fitted model:  
> # Find out what is our maximum sales:  
> indx <- which.max(df$sales)  
> # This will give you an index that we use to get that observation  
> df[indx,]  
      TV radio newspaper sales  
176 276.9  48.9      41.8    27
```

Linear regression: Analysis

```
Console ~/IIT/CS422/lectures/regression/ ↗  
> # Let's see how we predict using the fitted model:  
> # Find out what is our maximum sales:  
> indx <- which.max(df$sales)  
> # This will give you an index that we use to get that observation  
> df[indx,]  
      TV radio newspaper sales  
176 276.9  48.9      41.8    27  
> # What happens if we double our TV budget to $500K, add ~20% to our radio  
> # budget but keep the newspaper budget the same?  
> predict.lm(model, data.frame(TV=500, radio=58.68, newspaper=41.8),  
+           interval="confidence")  
      fit      lwr      upr  
1 36.84079 35.71472 37.96685
```

Asks R to calculate the CI
at 95% level (default).

Linear regression: Analysis

```
Console ~/IIT/CS422/lectures/regression/ ↗
> # Let's see how we predict using the fitted model:
> # Find out what is our maximum sales:
> indx <- which.max(df$sales)
> # This will give you an index that we use to get that observation
> df[indx,]
      TV radio newspaper sales
176 276.9  48.9      41.8    27
> # What happens if we double our TV budget to $500K, add ~20% to our radio
> # budget but keep the newspaper budget the same?
> predict.lm(model, data.frame(TV=500, radio=58.68, newspaper=41.8),
+           interval="confidence")
      fit      lwr      upr
1 36.84079 35.71472 37.96685
> # Hmm...we know from our analysis that newspaper is not a statistically
> # significant predictor. (Sad, it should be!) So, what if we zero out
> # the newspaper budget?
> predict.lm(model, data.frame(TV=500, radio=58.68, newspaper=0),
+           interval="confidence", level=0.99)
      fit      lwr      upr
1 36.88415 35.2061 38.56221
```

Note: You can specify the confidence level as a parameter

Linear regression: Analysis

```
Console ~/IIT/CS422/lectures/regression/ ↗
> # Let's see how we predict using the fitted model:
> # Find out what is our maximum sales:
> indx <- which.max(df$sales)
> # This will give you an index that we use to get that observation
> df[indx,]
      TV radio newspaper sales
176 276.9  48.9      41.8    27
> # What happens if we double our TV budget to $500K, add ~20% to our radio
> # budget but keep the newspaper budget the same?
> predict.lm(model, data.frame(TV=500, radio=58.68, newspaper=41.8),
+           interval="confidence")
      fit      lwr      upr
1 36.84079 35.71472 37.96685
> # Hmm...we know from our analysis that newspaper is not a statistically
> # significant predictor. (Sad, it should be!) So, what if we zero out
> # the newspaper budget?
> predict.lm(model, data.frame(TV=500, radio=58.68, newspaper=0),
+           interval="confidence", level=0.99)
      fit      lwr      upr
1 36.88415 35.2061 38.56221
```

To get prediction without an interval, you can simply say:

```
> res <- predict(model, newdata=x)
```

where x is a dataframe containing observations to be predicted.

Linear regression: Analysis

- Note:
 - **Confidence interval**: Tells you about the likely location of the true population parameter. For example, given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in each city, the 95 % confidence interval is [10985, 11528].
 - **Prediction interval** (in R, to get this, specify `interval="prediction"` parameter to `predict.lm()`): This interval must account for both the uncertainty in knowing the value of the true population parameter, plus the variance of the data. Given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in that city, the 95 % prediction interval is [7930, 14580].
 - Note that both intervals are centered at 11,256, but that the prediction interval is substantially wider than the confidence interval.

Linear regression: Analysis

- Factors that influence regression:
 - Multi-collinearity
 - In presence of multi-collinearity:
 - Cannot trust the learned coefficients (weights), thus making it tedious to assess the relative importance of predictors in explaining the variation caused to the response.
 - Standard errors of the affected coefficients are likely to be high.
 - Which also means that p-values for those coefficients become small, making it difficult to reject the null hypothesis.
- Resources:
 - Olsrr: Tools for building OLS regression models (see <https://olsrr.rsquaredacademy.com/index.html>).