Linear Regression Models (II)

Lecture outline

1. Assumptions of Linear Regression
2. R Squared and Adjusted R Squared
3. F-test for model significance
4. t-test for parameter significance

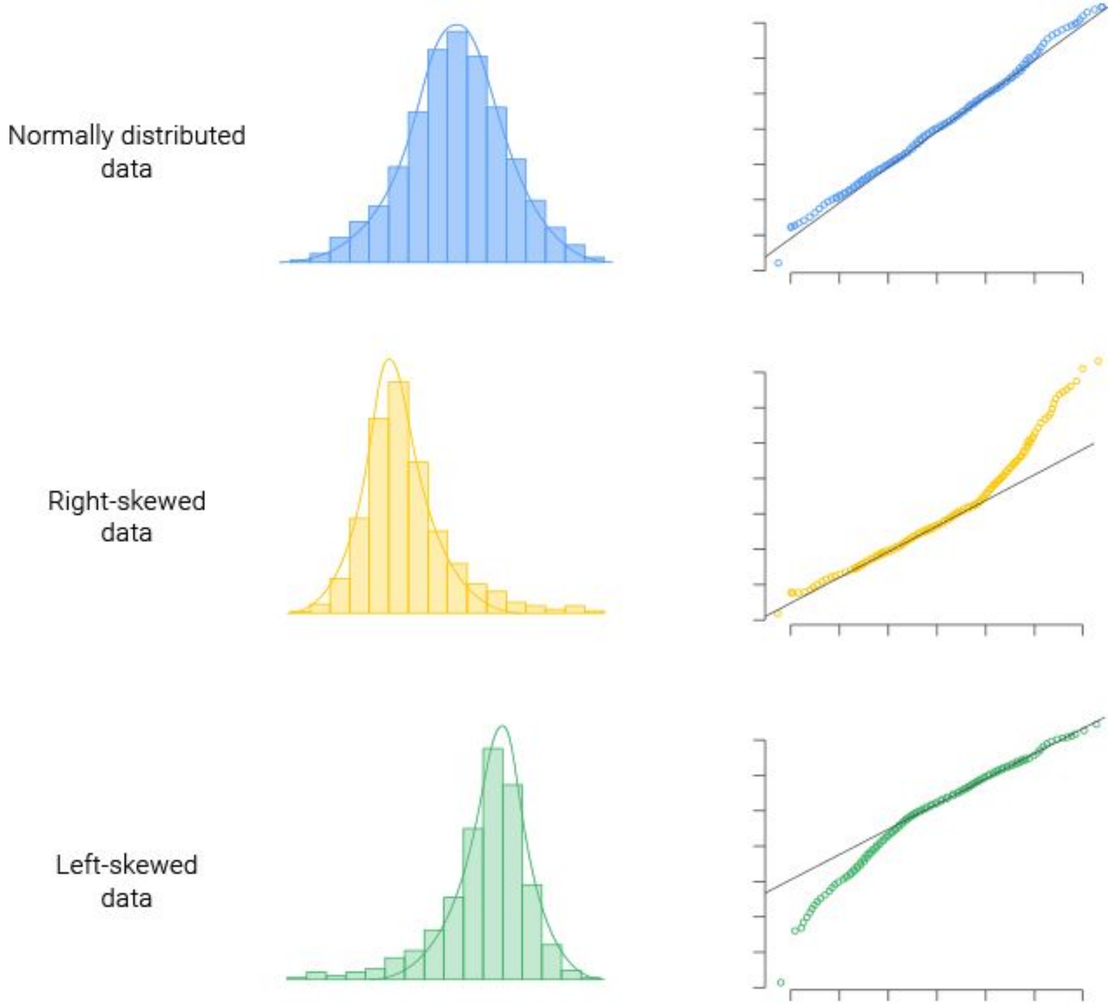
Assumptions of Linear Regression

- **Normality**: Multiple regression assumes that the error terms are normally distributed.
- **Linearity**: There must be linear relationship between response variable and independent variables (Scatterplots).
- **No Multicollinearity**: the independent variables are not highly correlated with each other (Correlation matrix).
- **Homoscedasticity**: the variance of error terms are similar across the values of the independent variables (Plot of residuals vs predictor variables).

Normality

Normality: Multiple regression assumes that the error terms are normally distributed.

Plot QQ (Quantile-quantile) plots are used to visually check the normality of the data.

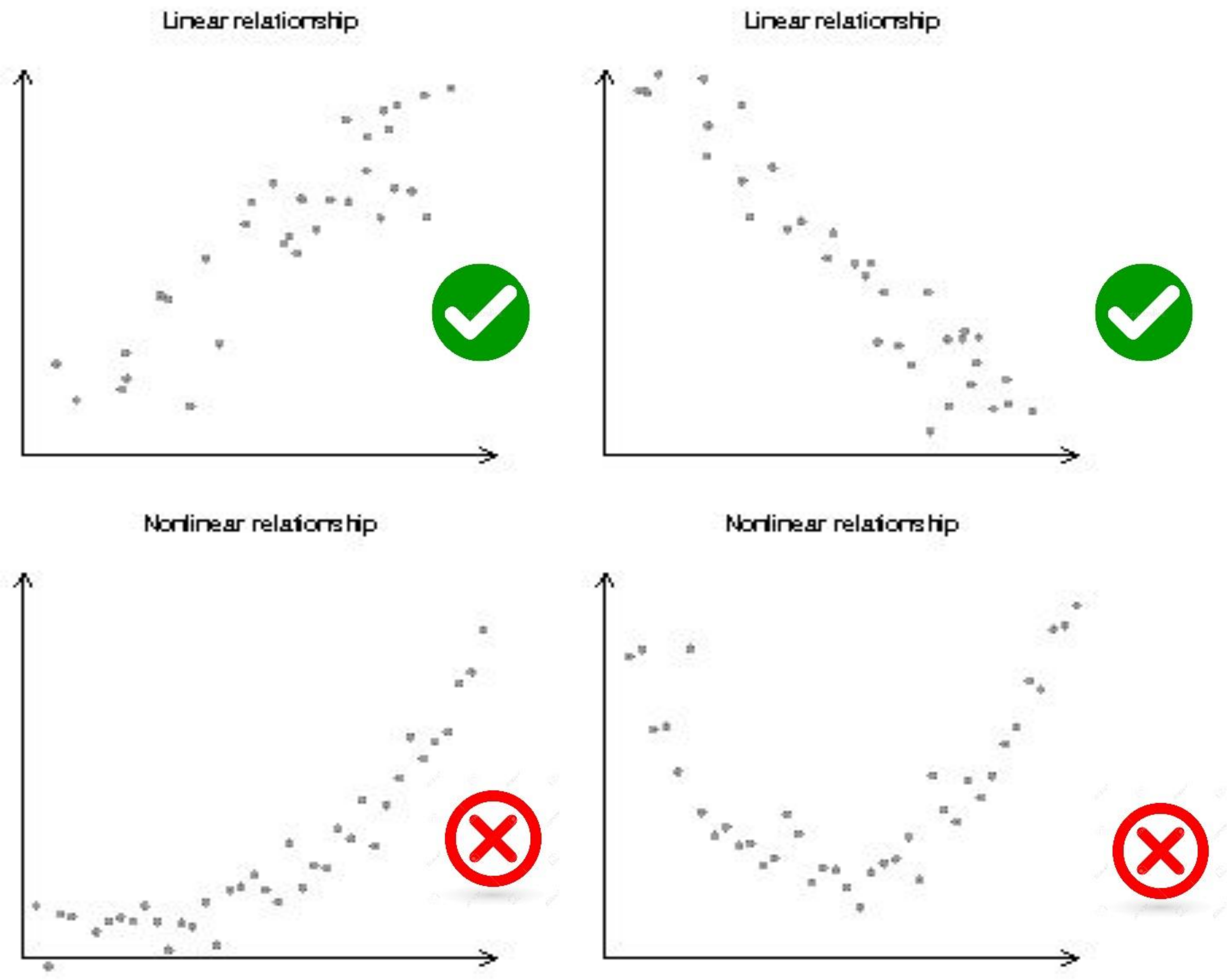


As all the points fall approximately along the straight line, we can assume normality.

R Syntax:
plot(model$residuals)

Linearity:

There must be linear relationship between response variable and independent variables.



No Multicollinearity

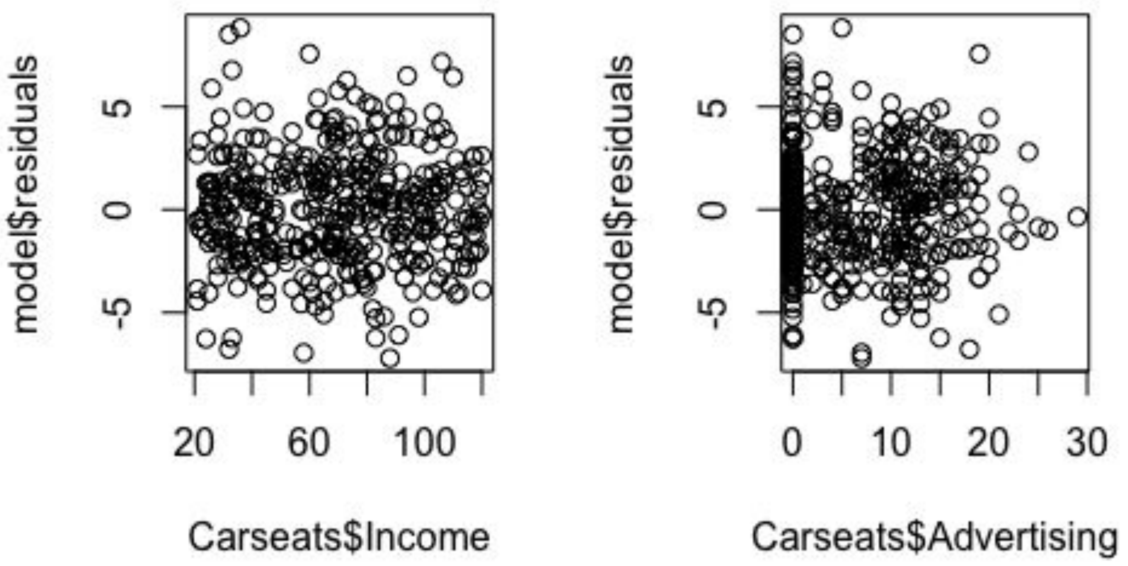
No Multicollinearity:

The independent variables are not highly correlated with each other.

	CompPrice	Income	Advertising
CompPrice	1.00	-0.08	-0.02
Income	-0.08	1.00	0.06
Advertising	-0.02	0.06	1.00

Homoscedasticity:

The variance of error terms are similar across the values of the independent variables (Plot of residuals vs predictor variables).




```
par(mfrow=c(1,2))
plot(Carseats$Income,model$residuals)
plot(Carseats$Advertising, model$residuals)
```

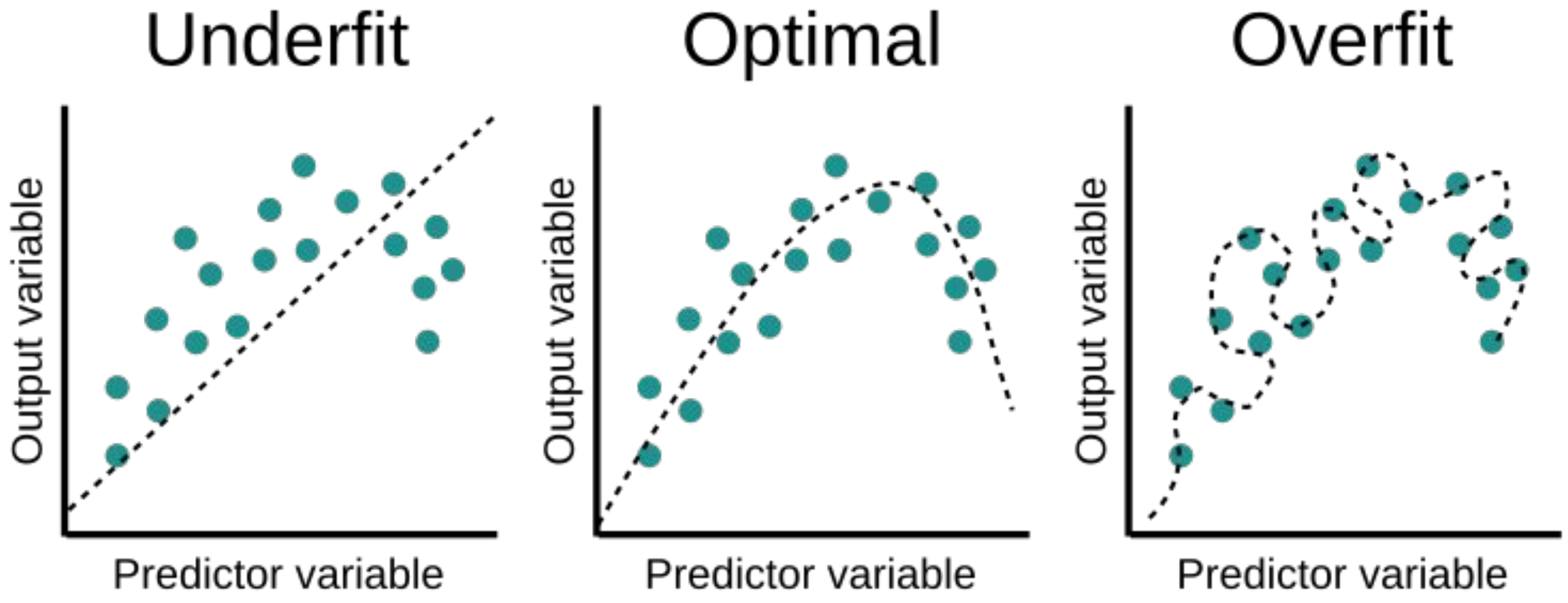
R-Squared

R-squared (R^2), also known as a Coefficient of Determination, is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.744554   0.390629  14.706  < 2e-16 ***
Income      0.013891   0.004843   2.868  0.00435 **
Advertising 0.111565   0.020384   5.473 7.86e-08 ***
UrbanNo     0.195994   0.296577   0.661  0.50909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.701 on 396 degrees of freedom
Multiple R-squared:  0.09221,    Adjusted R-squared:  0.08533
F-statistic: 13.41 on 3 and 396 DF,  p-value: 2.374e-08
```


Underfit
Optimal
Overfit
Output variable
Output variable
Output variable
Predictor variable
Predictor variable
Predictor variable


OVERFITTING

Adjusted R-Squared

- The **adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model.
- The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

$$\text{Adj } R^2 = 1 - \frac{\frac{SSE}{n-k}}{\frac{SST}{n-1}} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

Here n- # of observations

k - # of independent variables

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.744554   0.390629  14.706  < 2e-16 ***
Income      0.013891   0.004843   2.868  0.00435 **
Advertising 0.111565   0.020384   5.473 7.86e-08 ***
UrbanNo     0.195994   0.296577   0.661  0.50909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.701 on 396 degrees of freedom
Multiple R-squared:  0.09221,   Adjusted R-squared:  0.08533
F-statistic: 13.41 on 3 and 396 DF,  p-value: 2.374e-08
```

Testing for Significance: F-test

The *F* test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables.

The *F* test is referred to as the test for overall significance.

Testing for Significance: F-test

Hypotheses

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
H_a: At least one of the parameters (betas) is not equal to zero.

Test Statistics

F = MSR/MSE

Rejection Rule

Reject H_0 if p-value $\leq \alpha$ or if $F > F_\alpha$
where F_α is based on an F distribution
with k d.f. in the numerator and
$n - k - 1$ d.f. in the denominator.

Example

- *Let's use the insurance dataset to predict health care charges based on age, gender, bmi (body mass index) and smoker status.*

*Charges = β_0 + β_1*Age + β_2*Male + β_3*Bmi + β_4*Smoker + ε*

Hypotheses for Overall Significance (F-test):

Ho: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Ha: At least one of the betas is not zero.

Since p-value = 0.000…22 < α = 0.01,

We reject Ho and conclude that overall model is

Significant in predicting insurance charges.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11633.49     947.27 -12.281   <2e-16 ***
age            259.45      11.94  21.727   <2e-16 ***
sexmale       -109.04     334.66  -0.326    0.745
bmi            323.05      27.53  11.735   <2e-16 ***
smokeryes    23833.87     414.19  57.544   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6094 on 1333 degrees of freedom
Multiple R-squared:  0.7475,    Adjusted R-squared:  0.7467
F-statistic: 986.5 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Testing for Significance: t-test

Hypotheses

$H_0\text{:}\ \beta_i = 0$

$H_a\text{:}\ \beta_i \neq 0$

Test Statistics

$$t_{stat} = \frac{b_i}{s_{b_i}}$$

Rejection Rule

Reject H_0 if p-value $\leq \alpha$ or
if $|t| \geq t_{\alpha/2}$ where $t_{\alpha/2}$
is based on a t distribution
with $n - k - 1$ degrees of freedom.

Example

Let's use the insurance dataset to predict health care charges based on age, gender, bmi (body mass index) and smoker status.

$$Charges = \beta_0 + \beta_1{*}Age + \beta_2{*}Male + \beta_3{*}Bmi + \beta_4{*}Smoker + \varepsilon$$

Hypotheses for numerical variables for their significance (t-test): Age, Bmi

For Age:

Ho: $\beta_1 = 0$

Ha: $\beta_1 \neq 0$

For Bmi:

Ho: $\beta_3 = 0$

Ha: $\beta_3 \neq 0$

Since both p-values are smaller than alpha of 0.01, We reject Ho and conclude that age and bmi variables are significant.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11633.49     947.27 -12.281   <2e-16 ***
age            259.45      11.94  21.727   <2e-16 ***
sexmale       -109.04     334.66  -0.326    0.745
bmi            323.05      27.53  11.735   <2e-16 ***
smokeryes    23833.87     414.19  57.544   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6094 on 1333 degrees of freedom
Multiple R-squared:  0.7475,    Adjusted R-squared:  0.7467
F-statistic: 986.5 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Exercise 1

38 random movies were selected to develop a model for predicting their revenues.

We have the following variables in the dataset:

- ***USRevenue*** – movie’s revenue in the US (mln$)
- ***Rating*** – restrictions based on age (PG, PG-13, R)
- ***Budget*** – budget (expenditure) of the movie (mln$)
- ***Opening*** – revenue on the opening weekend (mln$)
- ***Theaters*** – number of theaters the movie was in for the opening weekend
- ***Opinion*** – IMDb rating (1 to 10, 10 being the best)

Exercise 1

a) Fit a model to predict USRevenue using Rating, Budget, Opening as independent variables.

b) Check the assumptions of the linear regression

c) Interpret the value of R^2 in part a.

d) Add Opinion as another explanatory variable into your model and see how your R^2 and adjusted R^2 changed. How would you explain this?

e) Test for overall significance of the model. Use $\alpha = 0.05$.

f) Test whether Budget, Opening and Opinion are significant variables separately. Use $\alpha = 0.05$.

THE END

Thank you for your attention!

Homework:

a) Refer to Housing Data. Fit a model to predict Price using all independent variables.

a) Check the assumptions of the linear regression

a) Interpret the value of R^2 in part a.

a) Test for overall significance of the model. Use $\alpha = 0.01$.

a) Test whether numerical variables are significant separately. Use $\alpha = 0.01$.

a) Now remove rooms and bathrooms from the model and compare with the original