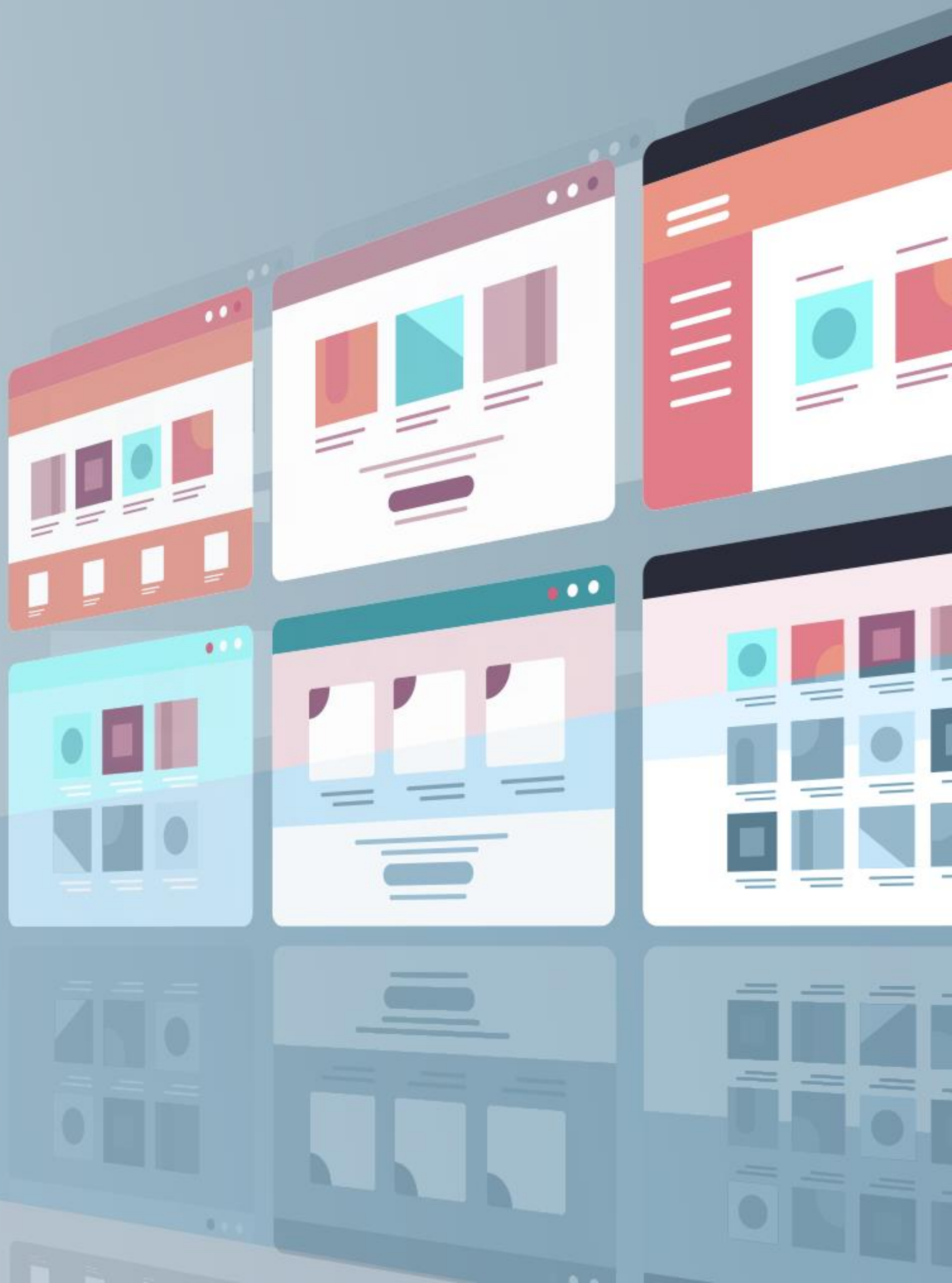
07

데이터분석방법론(1)

Multiple Regression

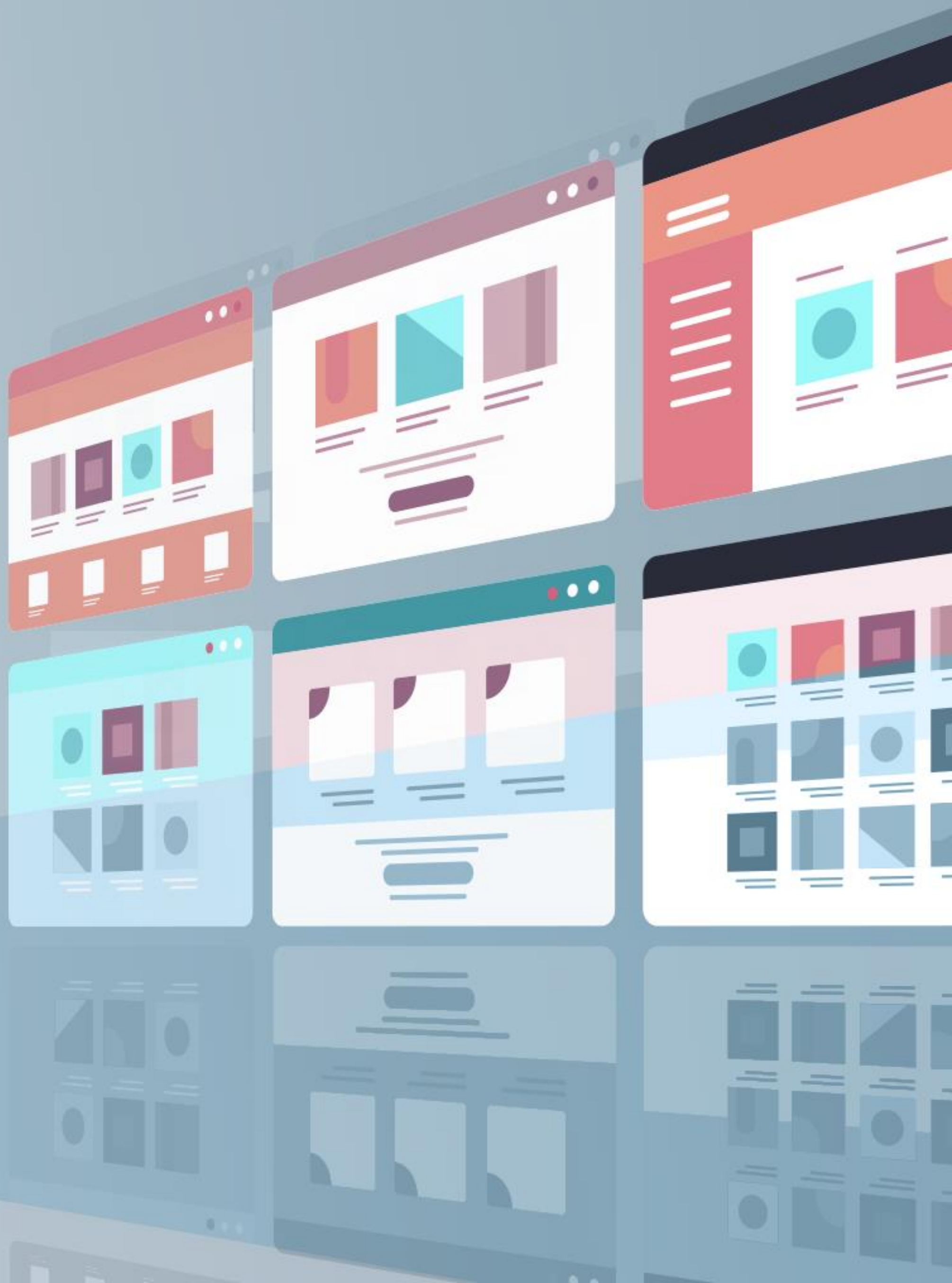
통계·데이터과학과 장영재 교수



학습목차

01

Introduction



1. Introduction

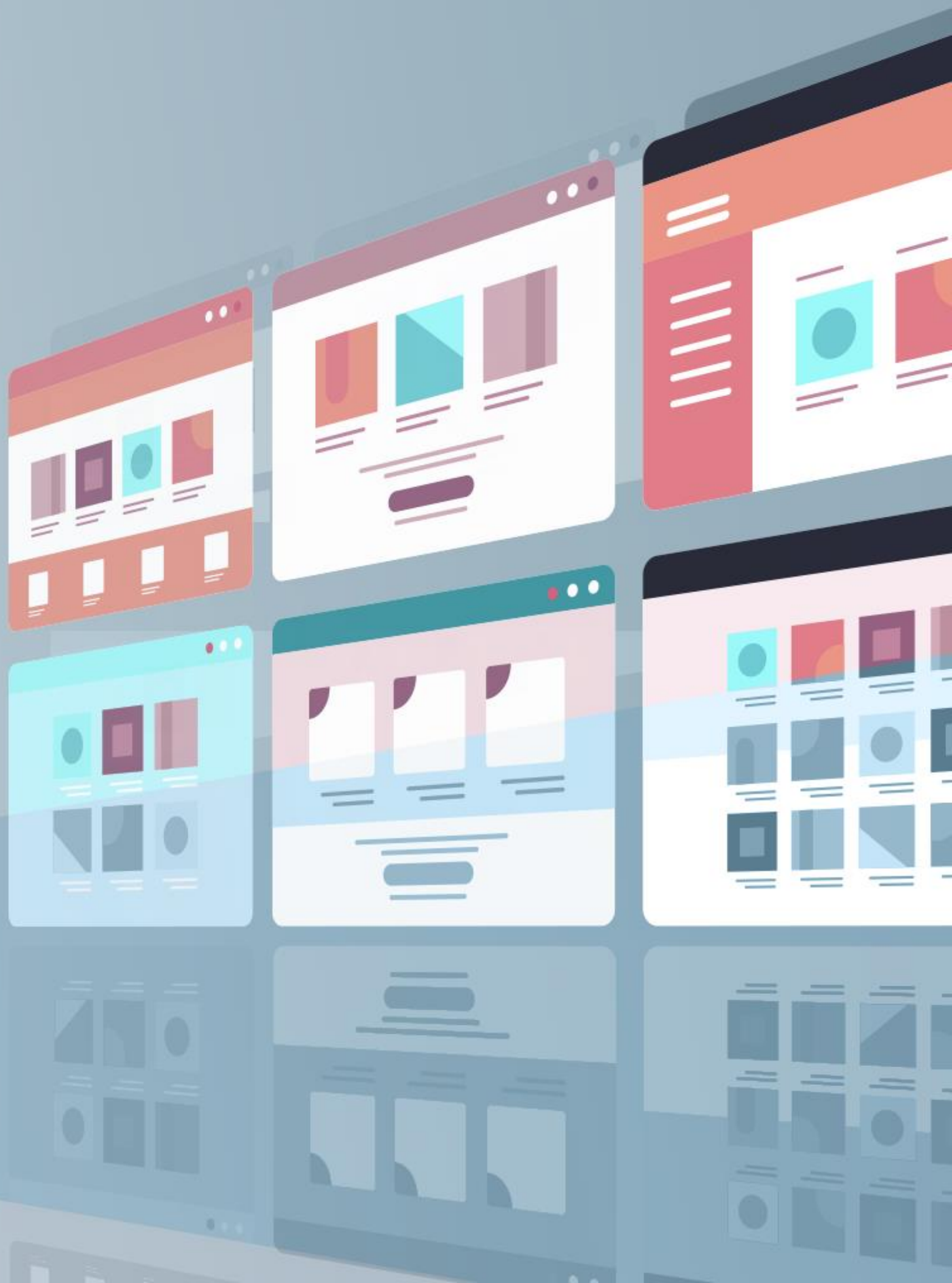
This lecture discusses the case of regression analysis with multiple predictors. The news is mainly the model search aspect, namely among a set of potential descriptive variables to look for a subset that describes the response sufficiently well. The basic model for multiple regression analysis is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

where $x_1, \cdots, x_k$ are explanatory variables (also called predictors) and the parameters $\beta_1, \cdots, \beta_k$ can be estimated using the method of least squares.

02

Model and Estimation



1. Linear Model

- One very general form for the model :

$$Y = f(X_1, X_2, X_3) + \varepsilon$$

where f is some unknown function and ε is an error

- Since we usually don't have enough data to try to estimate f directly, we usually have to assume that it has some more restricted form, perhaps linear as in

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- In a linear model the *parameters enter linearly* —the predictors do not have to be linear.

2. Matrix Representation

- Given the actual data, we may write:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

- Let

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$y = X\beta + \varepsilon$$

3. Least squares estimation

- Least square estimate of β, called $\widehat{\beta}$ minimizes SSE

$$\sum \varepsilon^2{}_i = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

$$\frac{\partial}{\partial \beta}(Y - X\beta)^T (Y - X\beta) = \frac{\partial}{\partial \beta}(Y^T - \beta^T X^T)(Y - X\beta)$$

$$= \frac{\partial}{\partial \beta}(Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta)$$

$$= \frac{\partial}{\partial \beta}(Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta)$$

$$= -2X^T Y + 2X^T X\beta = 0$$

3. Least squares estimation

- Least square estimate of β , called $\hat{\beta}$ minimizes

$$\sum \varepsilon^2_i = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

- Differentiating with respect to β and setting to zero, we find that $\hat{\beta}$ satisfies

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

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

- Predicted values : $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy,\quad H = X(X^T X)^{-1} X^T$

Hat Matrix

Residuals : $\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y} = (I - H)y$

Residual sum of squares : $\hat{\varepsilon}^T \hat{\varepsilon} = y^T (I - H)(I - H)y = y^T (I - H)y$

- Assume the errors are uncorrelated and have equal variance, $Var(\varepsilon) = I\sigma^2$

4. Mean and variance of $\widehat{\beta}$

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

- **Mean** $E\widehat{\beta} = (X^T X)^{-1} X^T X \beta = \beta$ **(unbiased)**

- $$\begin{aligned} \mathrm{var}(\widehat{\beta}) &= \mathrm{var}(Ay) \\ &= A\,\mathrm{var}(y)A^T \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\ &= (X^T X)^{-1} \sigma^2 \end{aligned}$$

- **Standard error of $\widehat{\boldsymbol{\beta}}_i$:** $se(\widehat{\beta}_i) = \sqrt{(X^T X)_{ii}^{-1} \widehat{\sigma}}$

5. Estimating σ^2

ANOVA Table

	SS	Df	MS	F-value
Regress Error	SSR SSE	P n-p-1	MSR MSE	MSR/MSE
Total	SST	n-1		

$$\hat{\sigma}^2 = SSE/(n-p-1) : MSE$$

Coefficient of determination : $R^2 = SSR/SST$

6. Example

```
> gfit = lm(Species ~ Area+Elevation+Nearest+Scruz+Adjacent, data=gala)
> summary(gfit)
 …
Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

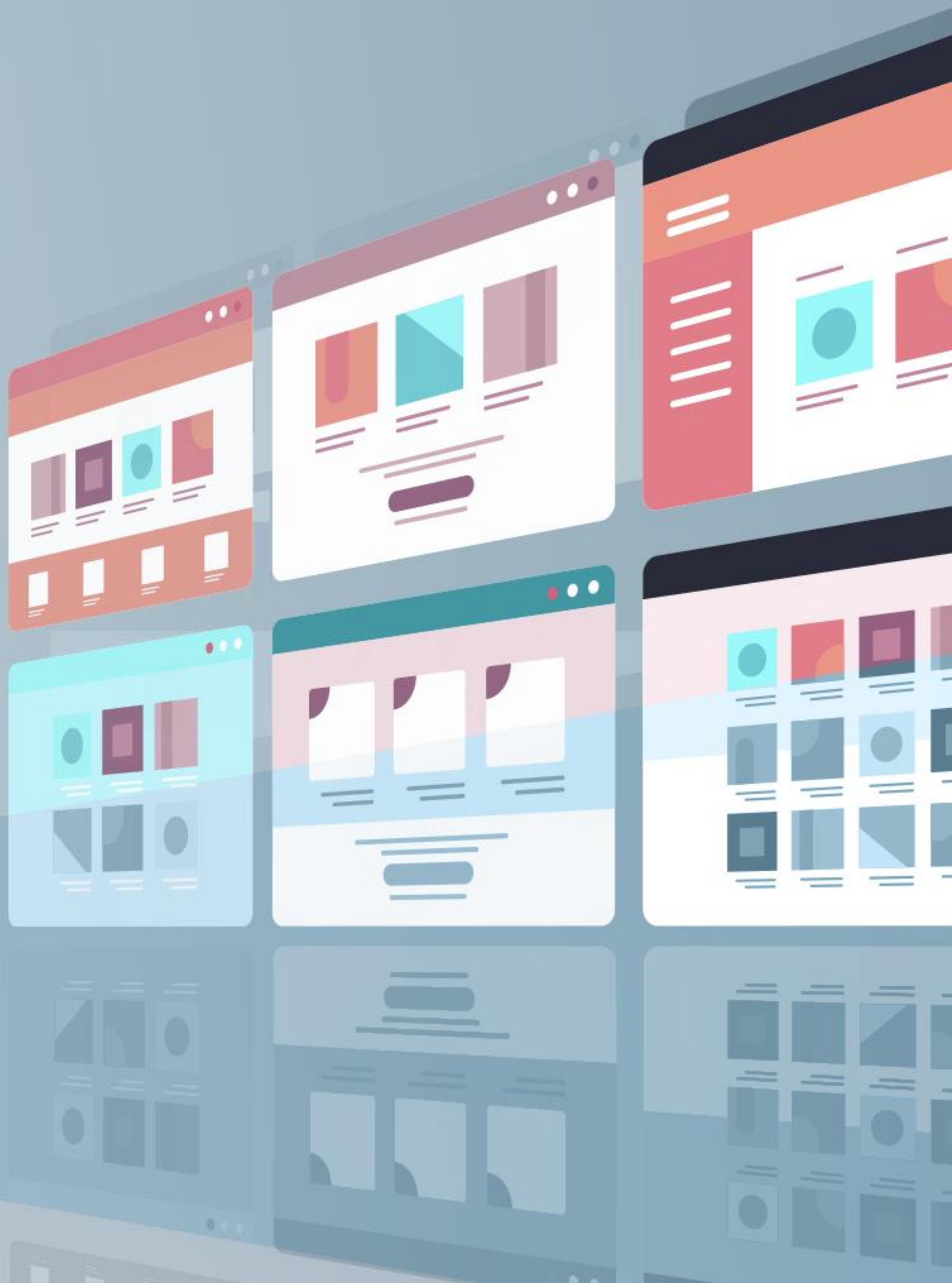
6. Example

```
> anova(gfit)
Analysis of Variance Table
Response: Species
          Df Sum Sq Mean Sq F value    Pr(>F)
Area       1 145470  145470 39.1262 1.826e-06 ***
Elevation  1  65664   65664 17.6613 0.0003155 ***
Nearest    1     29      29  0.0079 0.9300674
Scruz      1  14280   14280  3.8408 0.0617324 .
Adjacent   1  66406   66406 17.8609 0.0002971 ***
Residuals 24  89231    3718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

> names(gfit)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> gfit$coef
 (Intercept)         Area    Elevation      Nearest        Scruz     Adjacent
 7.068220709 -0.023938338  0.319464761  0.009143961 -0.240524230 -0.074804832
```

03

Inference : Example



1. Recall : The model

- Model

$$y = X\beta + \varepsilon$$

- We assume that the errors are independent and identically normally distributed with mean 0 and variance σ^2 , i.e.

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

$$y \sim N(X\beta, \sigma^2 I)$$

2. Examples

- Let's illustrate this test and others using an old economic dataset on 50 different countries. These data are averages over 1960-1970 (to remove business cycle or other short-term fluctuations). dpi is per-capita disposable income in U.S. dollars; ddpi is the percent rate of change in per capita disposable income; sr is aggregate personal saving divided by disposable income. The percentage population under 15 (pop15) and over 75 (pop75) are also recorded. The data come from Belsley, Kuh, and Welsch (1980).

```
> data(savings)
> head(savings, 3)
             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
Belgium   13.17 23.80  4.43 2108.47 3.82
```

3. Estimation and Hypothesis test

Test of all predictors

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,     Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

```
> 1-pf(5.756, 4,45)
[1] 0.0007900702
```

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ **Since the p-value is so small, this null hypothesis is rejected.**

3. Estimation and Hypothesis test

Testing just one predictor

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
```

- **Method 1 : using t value** $t_i = \hat{\beta}_i / se(\hat{\beta}_i)$
- **Method 2 : general F-testing approach**

3. Estimation and Hypothesis test

- **Method 2 : general F-testing approach**

```
> g2 <- lm(sr ~ pop75 + dpi + ddpi, data=savings)
> anova(g2,g)
Analysis of Variance Table

Model 1: sr ~ pop75 + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     46 797.72
2     45 650.71  1    147.01 10.167 0.002603 **
```

Understand that this test of pop15 is relative to the other predictors in the model, namely pop75, dpi and ddpi. If these other predictors were changed, the result of the test may be different. This means that it is not possible to look at the effect of pop15 in isolation.

4. Confidence intervals for prediction

- Given a new set of predictors, x_0 what is the predicted response? Easy — just $\hat{y}_0 = x_0^T \hat{\beta}$
- There are two kinds of predictions that can be made for a given x_0.

1. Suppose a new house comes on the market with characteristic x_0. Its selling price will be $x_0^T \hat{\beta} + \varepsilon$. Since $E\varepsilon = 0$, the predicted price is $x_0^T \hat{\beta}$ but in assessing the variance of this prediction, we must include the variance of ε.
2. Suppose we ask the question — "What would the house with characteristics x_0" sell for on average. This selling price is $x_0^T \hat{\beta}$ and is again predicted by $x_0^T \hat{\beta}$ but now only the variance in $\hat{\beta}$ needs to be taken into account.

- Most times, we will want the first case which is called "prediction of a future value" while the second case, called "prediction of the mean response" is less common.

4. Confidence intervals for prediction

- Now $\operatorname{var}(x_0{}^T \hat{\beta}) = x_0{}^T (X^T X)^{-1} x_0 \sigma^2$

- A future observation is predicted to be $x_0{}^T \hat{\beta} + \varepsilon$ (where we don't what the future ε will turn out to be).

So, $100(1-\alpha)\%$ confidence interval for a single future response is

$$\hat{y}_0 \pm t_{(\alpha/2,\phi)} \hat{\sigma} \sqrt{1 + x_0{}^T (X^T X)^{-1} x_0}$$

- If on the other hand, you want a confidence interval for the average of the responses for given x_0

$$\hat{y}_0 \pm t_{(\alpha/2,\phi)} \hat{\sigma} \sqrt{x_0{}^T (X^T X)^{-1} x_0}$$

4. Confidence intervals for prediction

- Suppose we want to predict the number of species (of tortoise) on an island with predictors 0.08,93,6.0,12.0,0.34(same order as in the dataset).
- Do it directly from the formula

```
> x0 <- c(1,0.08,93,6.0,12.0,0.34)
> y0 <- sum(x0*g$coef)
> y0
[1] 33.91967
> qt(0.975,24)
[1] 2.063899
> x <- cbind(1,gala[,3:7])
> x <- as.matrix(x)
> xtxi <- solve(t(x) %*% x)
> bm <- sqrt(x0 %*% xtxi %*% x0) *2.064 * 60.98
> bm
        [,1]
[1,] 32.89005
> c(y0-bm,y0+bm)
[1]  1.029614 66.809721
> bm2 <- sqrt(1+x0 %*% xtxi %*% x0) *2.064 * 60.98
> c(y0-bm2,y0+bm2)
[1] -96.16946 164.00879
```

$$\hat{y}_0 \pm t_{(\alpha/2,\phi)}\hat{\sigma}\sqrt{x_0{}^T(X^TX)^{-1}x_0}$$

$$\hat{y}_0 \pm t_{(\alpha/2,\phi)}\hat{\sigma}\sqrt{1+x_0{}^T(X^TX)^{-1}x_0}$$

4. Confidence intervals for prediction

- There is a more direct method for computing the CI. The function predict() requires that its second argument be a data frame with variables named in the same way as the original dataset: |

```
> new = data.frame(Area=0.08,Elevation=93,Nearest=6.0,Scruz=12,Adjacent=0.34)
> predict(g, new, interval="confidence")
       fit      lwr      upr
1 33.91967 1.033826 66.80551
> predict(g, new, interval="prediction")
       fit      lwr      upr
1 33.91967 -96.1528 163.9921
```

$$\hat{y}_0 \pm t_{(\alpha/2,\phi)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

$$\hat{y}_0 \pm t_{(\alpha/2,\phi)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

5. Regression Diagnostics

- After establishing a regression model and performing estimation and testing of coefficients, it is necessary to review in detail whether the fitted model is stable and whether the assumptions are reasonable.

1. Review whether assumptions are violated through residual analysis
2. Detection of outliers or influential points
3. Review the stability of the model by examining the correlation between independent variables

다음시간 안내

09

Unusual and Influential Data