Lecture 14: Linear Regression

Outline

- How to fit a linear regression model
- How to evaluate the regression model
- Standard error of regression estimates
- Coefficient of determination

Introduction to linear regression

- The Pearson correlation measures the degree to which a set of data points form a straight line relationship.
- **Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.

We find correlation on a scatter plot;
Now,
We want to “draw” a line to capture their relationship.

Person	Family Income (in $1000)	Student's Average Grade
A	31	72
B	38	86
C	42	81
D	44	78
E	49	85
F	56	80
G	58	91
H	65	89
I	70	94
J	90	83
K	92	90
L	106	97
M	135	89
N	174	95

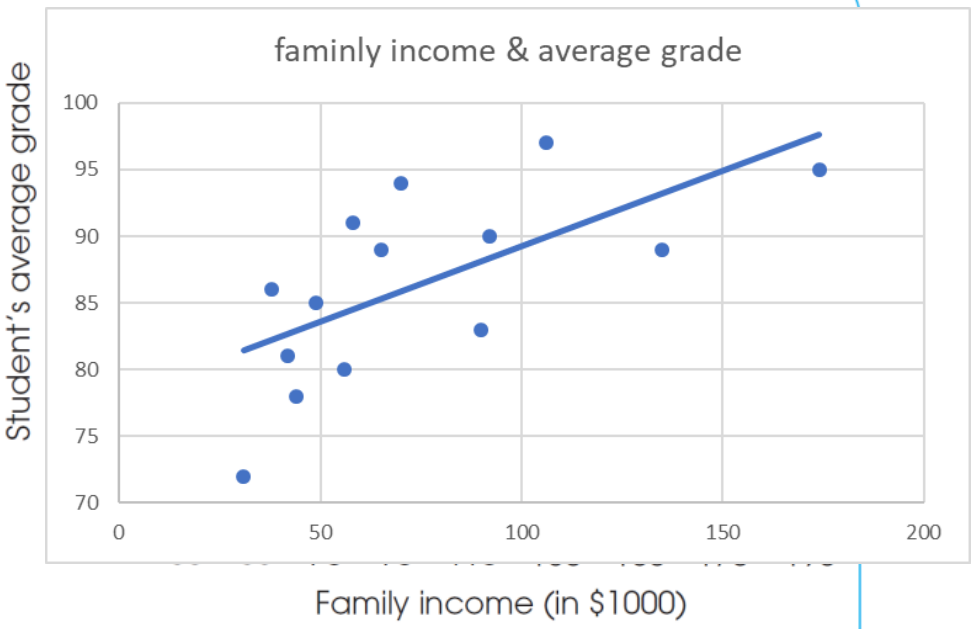


FIGURE 14.1

Correlational data showing the relationship between family income (X) and student grades (Y) for a sample of $n = 14$ high school students. The scores are listed in order from lowest to highest family income and are shown in a scatter plot.

Introduction to linear regression (cont.)

- Any straight line can be represented by an equation of the form $Y = bX + a$, where b and a are constants.
- The value of b is called the slope constant and determines the direction and degree to which the line is tilted.
- The value of a is called the Y-intercept and determines the point where the line crosses the Y-axis.

Linear Regression

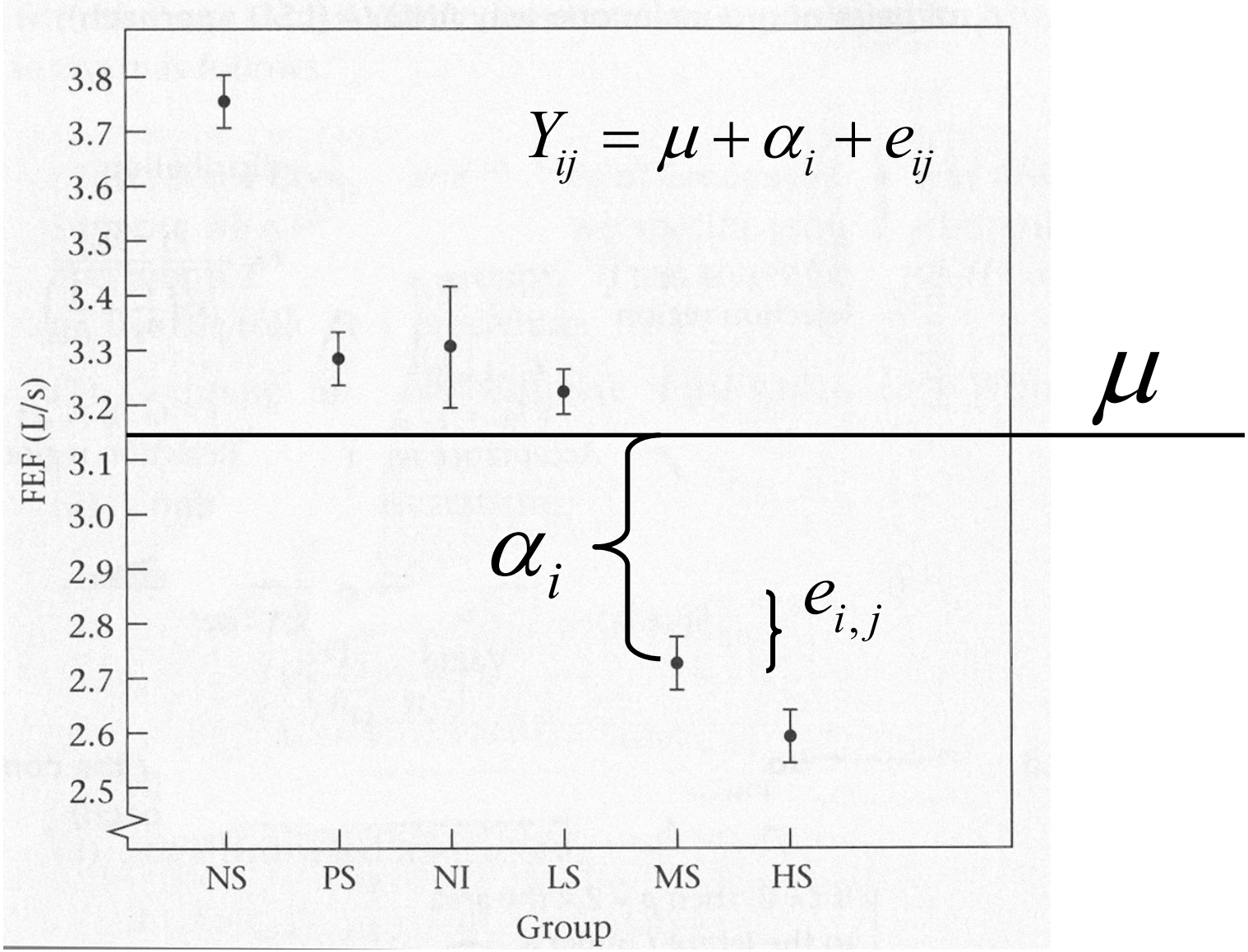
- Y is referred to as the dependent variable; X is referred to as the independent variable.

- A set of data consists of n pairs of observations, denoted (x_1,y_1), (x_2, y_2), …, (x_n, y_n)

- Assume: $y= \alpha+ \beta x +e$, where $e \sim N(0, \sigma^2)$
=> each y_i is an independent observation from the distribution $N(\alpha+\beta x_i, \sigma^2)$

- This means that $E(Y_i|X_i)= \alpha+\beta X_i$; To characterize the model, we need estimates of the parameters, α β, and σ^2.

ANOVA is also a linear model



Regression analysis

- Assess the association: Look at the distribution of **one variable**

-- the response variable, outcome variable, or dependent variable （因变量）

- as **another variable** is varied

-- the explanatory variable, covariate, predictor, or independent variable （自变量）

- The dependent variable is modeled as a **function** of the independent variables, corresponding parameters, and an error term.

-- regression equation（回归方程）: $Y = f(X, \beta)$

- The parameters are estimated so as to give a **"best fit" of the data** （数据拟合）
- As opposed to correlation, this is **not symmetric:** we want to use X to predict Y (not Y to predict X)

Fit the regression line

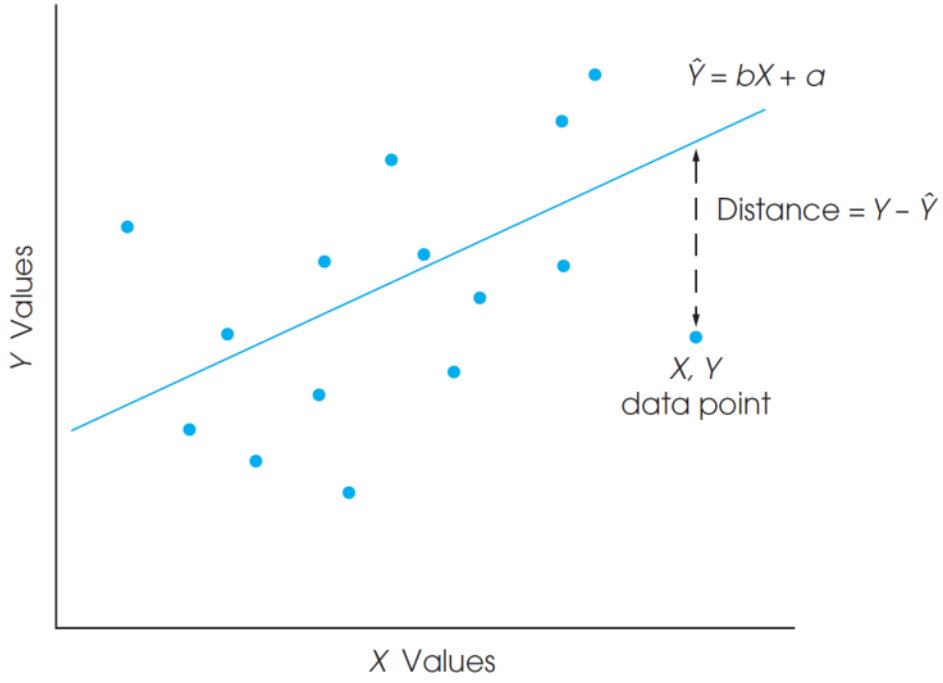
Let d_i be the deviation of y_i (actual data) from the point on the line $\hat{y}$ (i.e.the predicted value).

$$
\begin{aligned}
d_i &= y_i - \hat{y}_i \\
&= y_i - (\hat{\alpha} + \hat{\beta} x_i) \\
&= y_i - \hat{\alpha} - \hat{\beta} x_i \\
&= y_i - a - b x_i
\end{aligned}
$$

We want to make these distances d_i's as small as possible.

FIGURE 14.15

The distance between the actual data point (Y) and the predicted point on the line ($\hat{Y}$) is defined as $Y - \hat{Y}$. The goal of regression is to find the equation for the line that minimizes these distances.

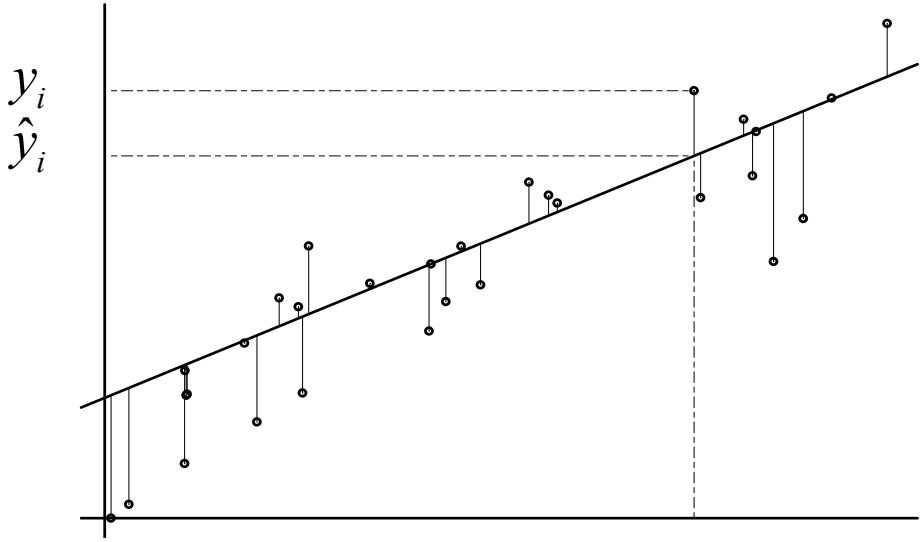


Least Squares for Fitting the regression line

Let S be **the sum of squared deviations** of points from the line in the y direction :

$$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

The least squares line is the line y = a+bx that minimizes S.



Least Squares for Fitting the regression line

By solving $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$, we can get that the values

for α and β that minimize S are given by:

$$\hat{\alpha} = a = \bar{y} - b\bar{x},$$

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\overline{x^2}} = \frac{\operatorname{cov}(x, y)}{SS_x} * \mathrm{n}$$

Prove it by yourself using the hint: The point $\left(\bar{x}, \bar{y}\right)$ is always on the line

Some Notations

$$SP = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum XY - \frac{\sum X \sum Y}{n} = cov(X,Y) * \mathrm{n}$$

$$SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\hat{\beta} = b = SP / SS_x$$

Introduction to linear regression (cont.)

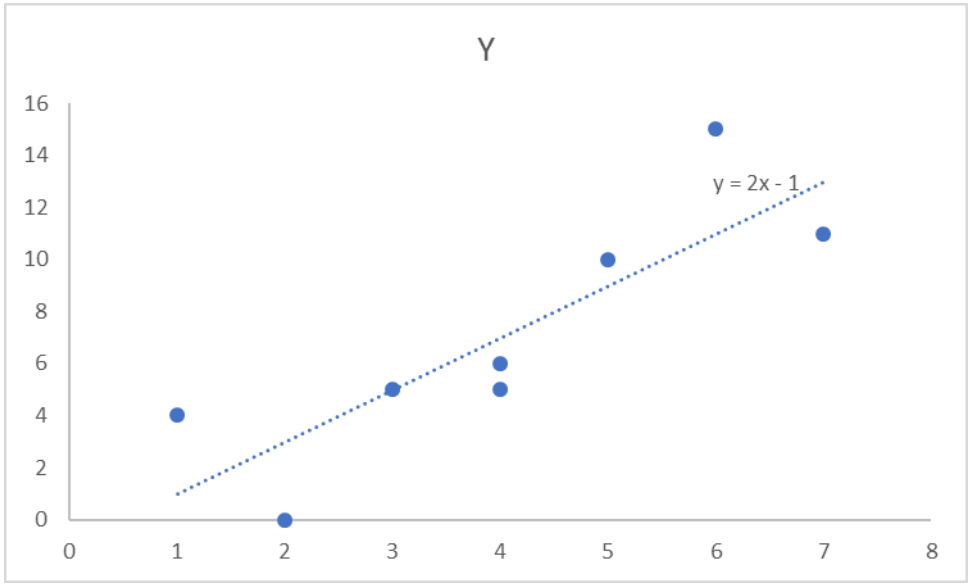
The equation for the regression line is

$$\hat{Y} = bX + a$$

Where $b = \frac{SP}{SS_X}$ and $a = \bar{Y} - \text{b}\,\bar{X}$

The scores in the following table are used to demonstrate the calculation and use of the regression equation for predicting Y.

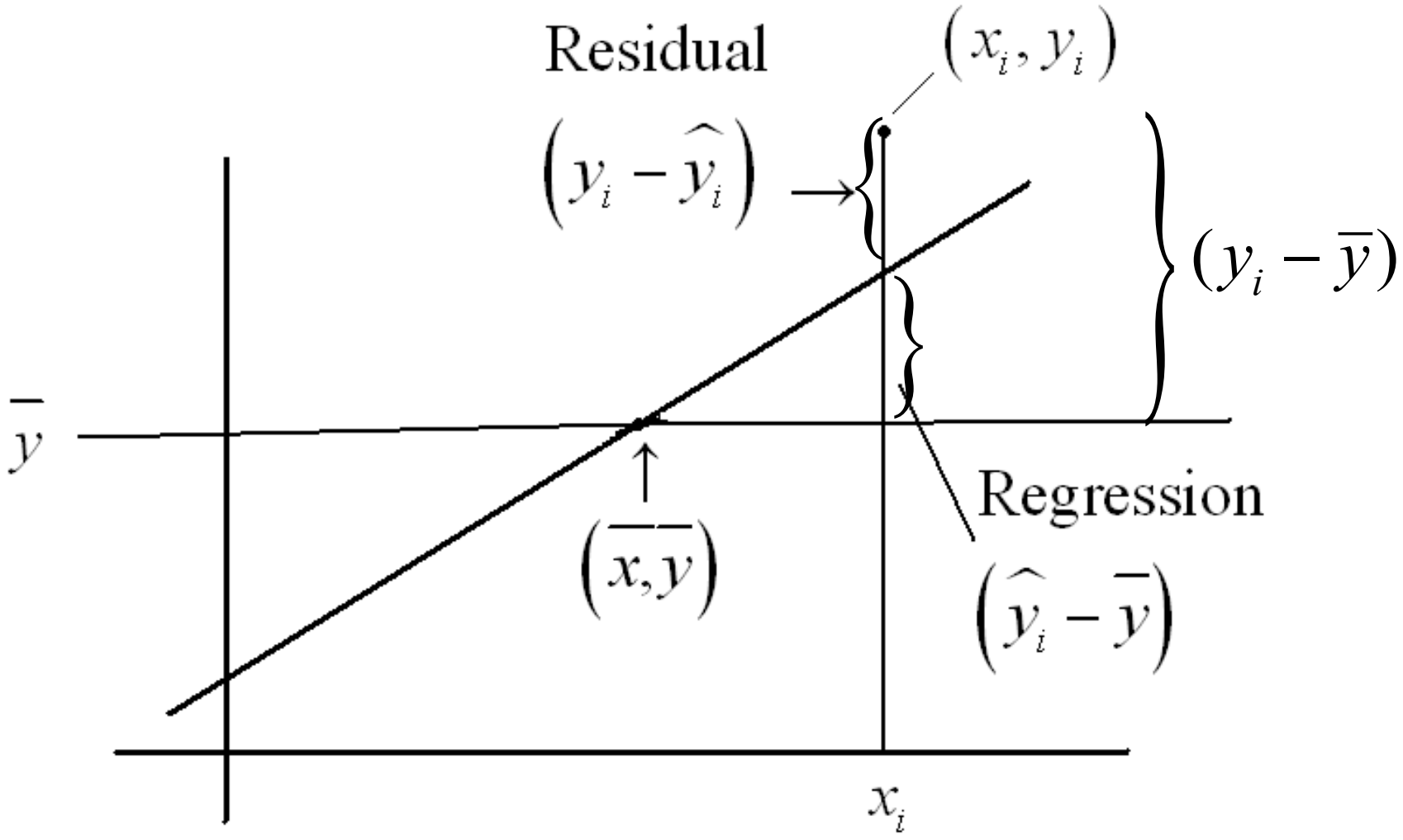
X	Y	$X - M_X$	$Y - M_Y$	$(X - M_X)^2$	$(Y - M_Y)^2$	$(X - M_X)(Y - M_Y)$
5	10	1	3	1	9	3
1	4	–3	–3	9	9	9
4	5	0	–2	0	4	0
7	11	3	4	9	16	12
6	15	2	8	4	64	16
4	6	0	–1	0	1	0
3	5	–1	–2	1	4	2
2	0	–2	–7	4	49	14
				$SS_X = 28$	$SS_Y = 156$	$SP = 56$



b = SP/SSx
= 56/28 = 2

a = $\bar{Y}$ – b* $\bar{X}$
= 7 – 2 * 4
= -1

Hypothesis testing for regression: Partitioning the Sums of Squares



$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Deviation from the mean *Residual Component* *Regression Component*

For the regression model: The smaller the residual the better the model

SS_y

Total sum of squares: $SS_{TOT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Regression sum of squares: $SS_{REG} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Residual sum of squares: $SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{y})^2$

$$SS_{TOT} = SS_{REG} + SS_{RES}$$

Also called SS_{error}

Coefficient of Determination: r^2

$$r^2 = {SS_{REG}}/{SS_{TOT}} = {(SP^2/SS_X)}/{SS_Y}$$

Thus, r^2 represents the proportion of the variance (of Y) that can be explained by the variable X.

For simple linear regression $y \sim \alpha + \beta x$, r^2 is simply the square of the sample correlation coefficient (r).

r^2=1: all variation in Y explained by variation in X.
r^2=0: X gives no information on Y.

Variance explained is a very important concept in modeling, indicating how good is your model in explaining your data.


SS_Y
$SS_{regression}$
r^2SS_Y
$SS_{residual}$
$(1 - r^2)SS_Y$
$df_Y = n - 1$
$df_{regression} = 1$
$df_{residual} = n - 2$

Introduction to Linear Regression (cont.)

- The ability of the regression equation to accurately predict the Y values is measured by first computing the proportion of the Y-score variability that is predicted by the regression equation and the proportion that is not predicted.

$$\text{Predicted variability} = SS_{\text{regression}} = r^2 SS_Y \qquad \text{with } df = 1$$

$$\text{Unpredicted variability} = SS_{\text{resicual}} = (1 - r^2) SS_Y = \Sigma(Y - \hat{Y})^2 \qquad \text{with } df = n - 2$$

Introduction to Linear Regression (cont.)

- Finally, the overall significance of the regression equation can be evaluated by computing an F-ratio.
- A significant F-ratio indicates that the equation predicts a significant portion of the variability in the Y scores (more than would be expected by chance alone).
- To compute the F-ratio, you first calculate a variance or MS for the predicted variability and for the unpredicted variability:

Introduction to Linear Regression (cont.)

$$MS_{regressiion} = \frac{SS_{regression}}{df_{regression}} \quad \text{and} \quad MS_{residuals} = \frac{SS_{residuals}}{df_{residuals}}$$

The F-ratio is

$$F = \frac{MS_{regressiion}}{MS_{residuals}} \quad \text{with } df = 1, n - 2$$

Using the same example above

Regression line: $Y = 2X - 1$

$SS_{REG} = SP^2/SS_x = (56)^2/28 = 112$

$SS_{TOT} = SS_y = 156$

$\mathbf{r^2 =}$ 112/156 = 0.718

This means the linear regression model can explain 71.8% of the variance in the data.

Relationship between the regression coefficient (b) and the correlation coefficient (r).

$$b = \frac{SP}{SS_x}$$

$$r = \frac{SP}{\sqrt{SS_x SS_y}}$$

$$b = r\sqrt{\frac{SS_y}{SS_x}}$$

这里，虽然公式和运算联系在一起了，但是相关和回归还是含义不同！

Relationship between the regression coefficient and the standard error

$$r^2 = {SS_{REG}}/{SS_{TOT}} = {(SS_{TOT} - SS_{RES})}/{SS_{TOT}}$$

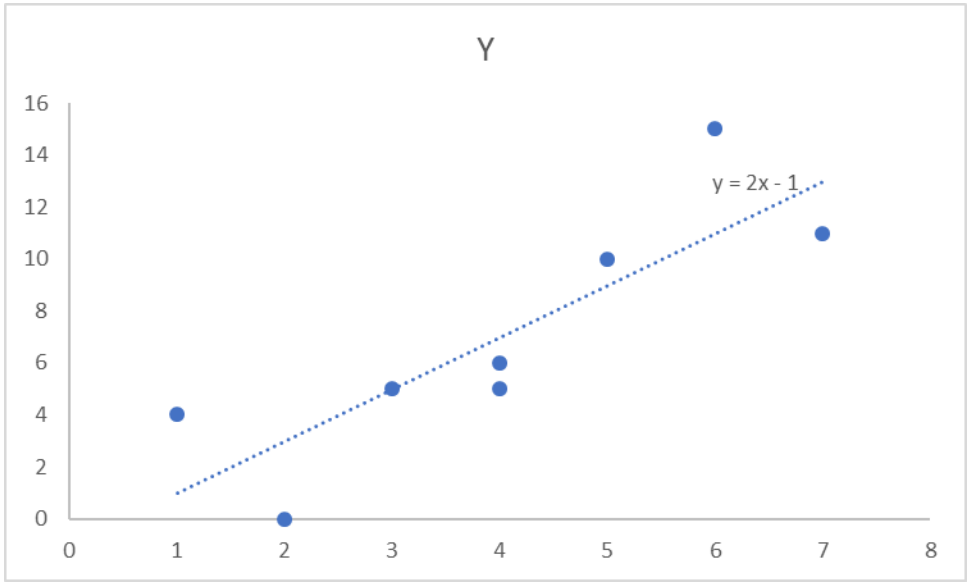
$$= {(SS_Y - SS_{error})}/{SS_Y} \longrightarrow SS_{error} = (1 - r^2)SS_y$$

The Standard Error of the estimates (predictions)

We want to know the variability of the estimates from the linear regression model. $SS_{residue}$ (SS_{error} in the textbook) shows the overall error that the linear regression model produces.

$$SS_{error} = SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$MS_{error} = SS_{error} / df$$ **df = n-2**



MS_{error} reflects the average error the model makes.

估计的标准误是$\sqrt{MSE}$

Standard error of estimate

- The unpredicted variability can be used to compute the standard error of estimate which is a measure of the average distance between the actual Y values and the predicted Y values.

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{df}}$$

Assumptions for regression

- **Assumptions (similar to Pearson correlation)**:
  - Both X and Y should be continuous variables (interval or ratio data)
  - The regression residues should follow a normal distribution
  - Each observation (a pair of x and y) should be independent. In other words, no related measures are allowed.
  - Homoscedasticity: the variance of residues should be the same across different levels of independent variable (x). This assumption is not mandatory.

Requirements:

- *No outliers*
- *The effect is indeed linear*

Usages of Regression Analysis

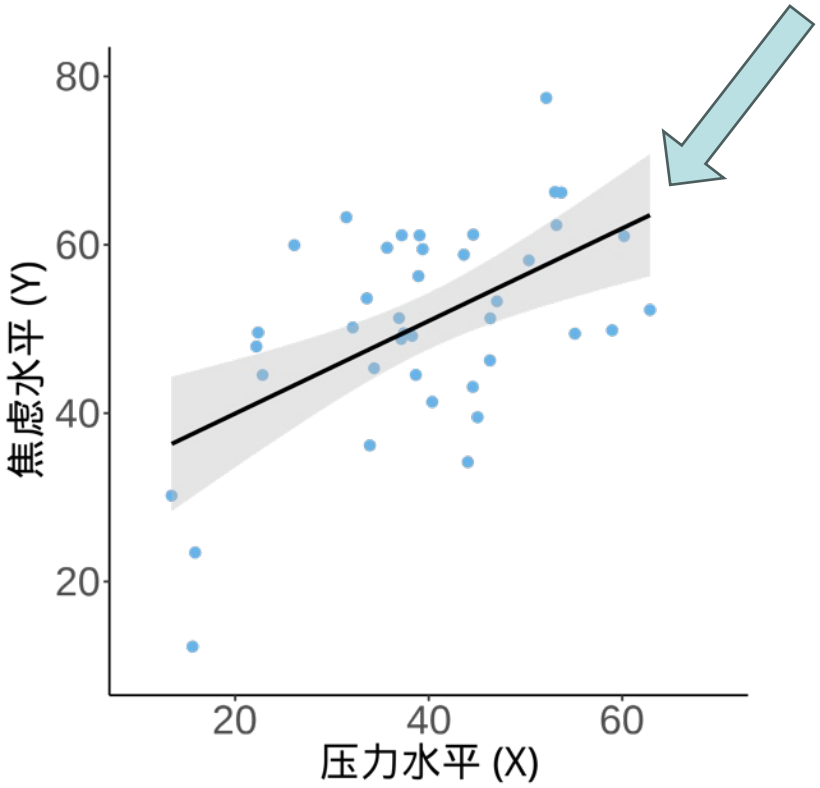
1) **Modeling & Inference**: Provides numerical measures of the relationship between the two variables.

2) **Prediction**: Allows prediction of the values of one variable when the value of the other variable is known. This prediction is not with certainty but we can say something about the mean value and the variability of the predicted value.

3) **Hypothesis Testing:** Allows assessment of significance of direction of trend.

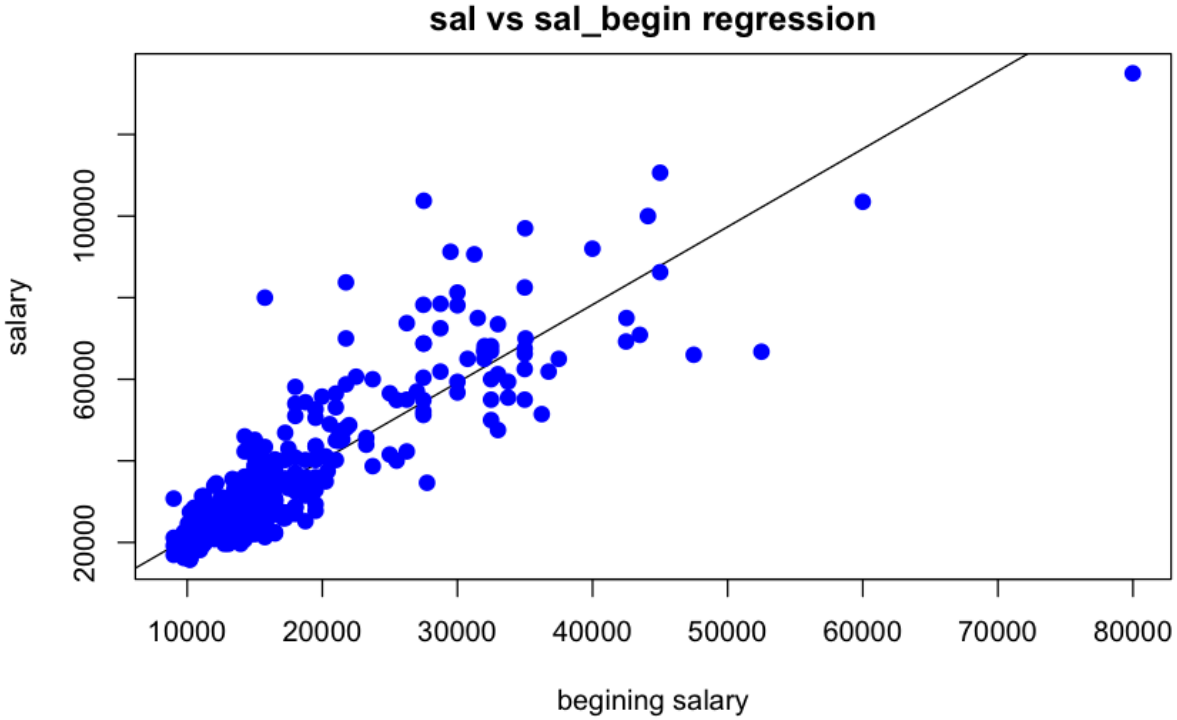
Things not covered but may be useful

- Hypothesis testing (whether b=0??) and confidence interval for parameters.
- Confidence interval of a specific prediction (y_hat).



R demo

Using beginning salary to predict the current salary




```
plot(x,y,col = "blue",main = "sal vs sal_begin regression",
abline(lm(y~x)),cex = 1.3,pch = 16,xlab = "begining salary",ylab = "salary")
```

```
Call:
lm(formula = salary ~ salbegin, data = reg_data)

Residuals:
   Min     1Q Median     3Q    Max
-35424  -4031  -1154   2584  49293

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.928e+03  8.887e+02    2.17   0.0305 *
salbegin    1.909e+00  4.741e-02   40.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8115 on 472 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7741
F-statistic:  1622 on 1 and 472 DF,  p-value: < 2.2e-16
```

- a and b
- Their SEM
- Their t tests

- SEM
- R^2
- F test

Coefficient of determination and goodness of fit

1. **r^2** , coefficient of determination（决定系数）：回归模型可以解释的方差的%

$$r^2 = {(SS_Y - SS_{error})}/{SS_Y} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

2. **Adjusted r^2** , 校正的r square：修正回归模型中数据量(n)和模型复杂程度（模型中的自由参数的数目）的影响

$$Adjusted\ r^2 = 1 - \frac{\sum(y-\hat{y})^2/(n-k-1)}{\sum(y-\bar{y})^2/(n-1)}$$

F test for the model

```
Residual standard error: 8115 on 472 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7741
F-statistic:  1622 on 1 and 472 DF,  p-value: < 2.2e-16
```

- Test whether the model is significant, or whether the X is important for predicting Y.
- Using the analogy of ANOVA, whether the MS of the model is significant.

$$F = \frac{\text{回归均方}}{\text{残差均方}} = \frac{MS_{regression}}{MS_{error}} = \frac{\sum(\hat{y} - \bar{y})^2/p}{\sum(y - \hat{y})^2/(n - p - 1)}$$

- P is the number of predictors used in the model. Here the number of X is only 1.

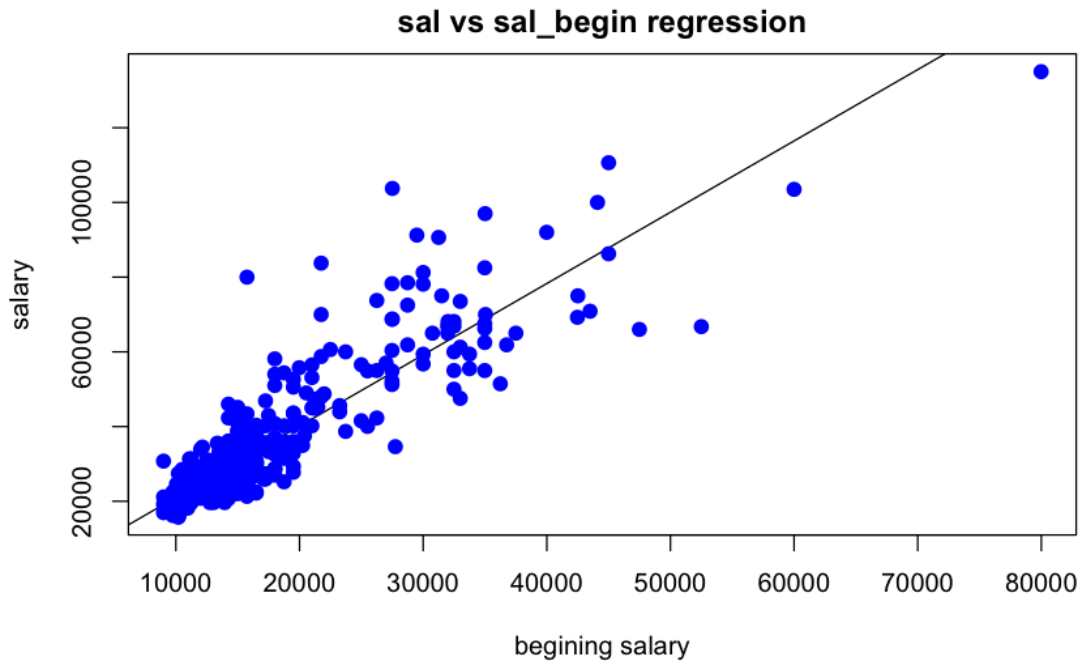
T tests for each regression coefficient

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.928e+03  8.887e+02    2.17   0.0305 *
salbegin    1.909e+00  4.741e-02   40.28   <2e-16 ***
```

For a: $t = \frac{a}{SE_a} = \frac{\beta_0}{SE_a}$ df = 1, n-p-1

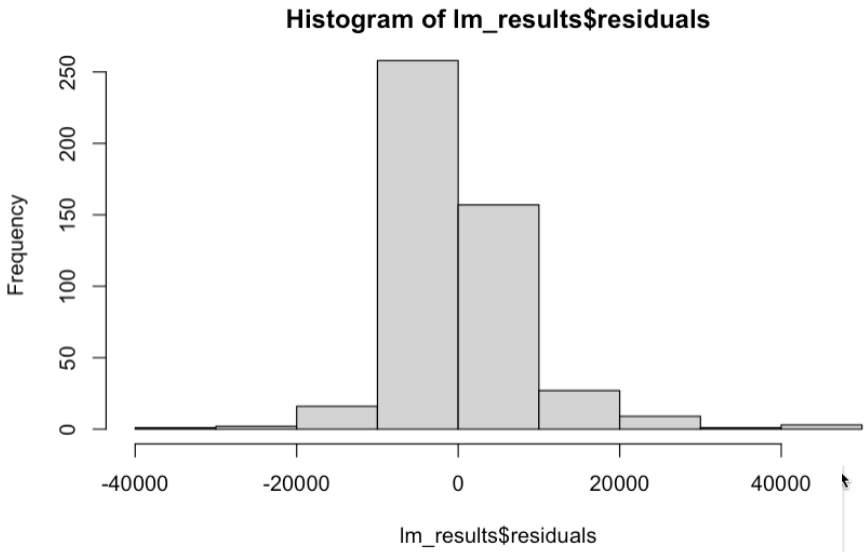
For b: $t = \frac{b}{SE_b} = \frac{\beta_1}{SE_b}$ df = 1, n-p-1

Check the predictions

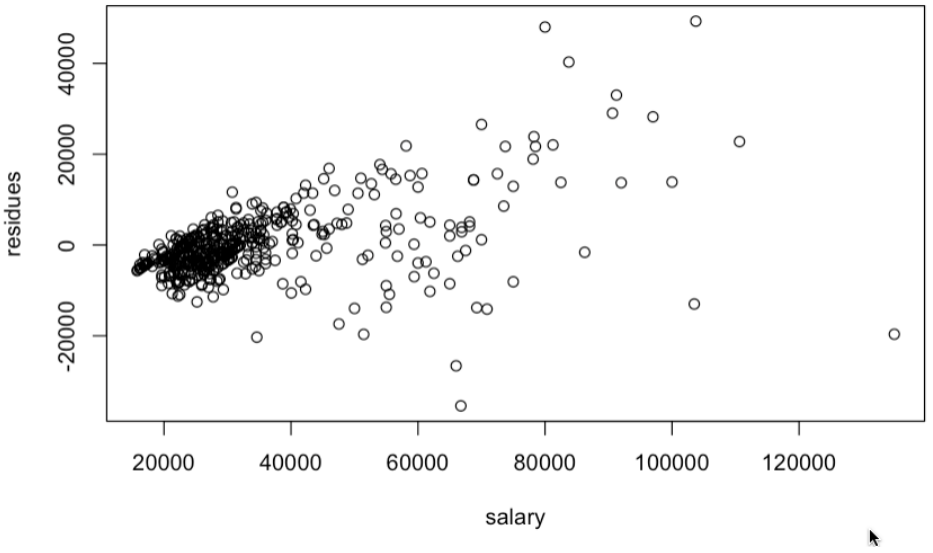


plot(y,lm_results$fitted.values,xlab='salary',ylab='predicted salary')

Check the residues

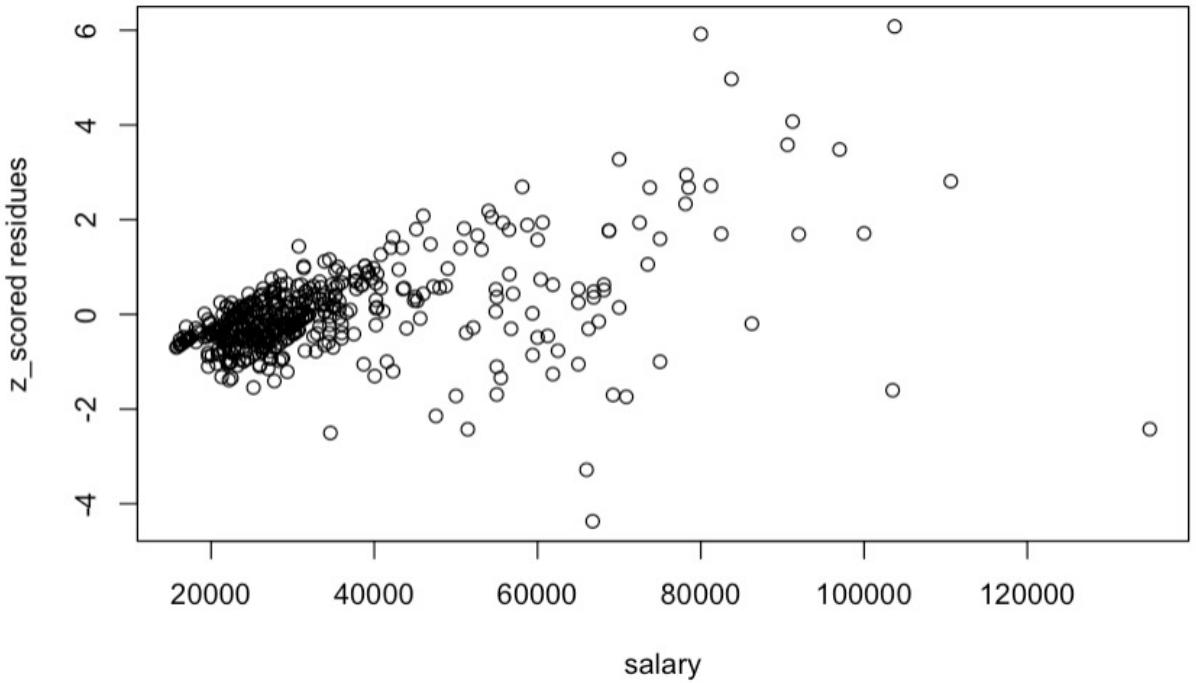


- Check whether the residues are symmetrical around 0 like a Gaussian
- Check whether the residues does not depend on the actual data




```
hist(lm_results$residuals)
plot(y,lm_results$residuals,xlab='salary',ylab='residues')
```

Check the residues



Residues larger than 2*SEM are possible outliers, which have large influence on the fitting.

```
plot(y,scale(lm_results$residuals),xlab='salary',ylab='z_scored residues')
```