Chapter 15: Linear Regression II

DSCC 462

Computational Introduction to Statistics

Anson Kahng

Fall 2022

Plan For Today

- Learn to evaluate how good our linear regression is
- Introduce multiple regression (and inference for multiple regression)
- Learn how to include indicator variables and allow for interactions between variables

Evaluating Model Fit

- Once we fit a regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, we must then determine how well this line actually fits our data
- Numerical and graphical evaluations of model fit:
    - Coefficient of determination (R^2)
    - Residual plots

Coefficient of Determination

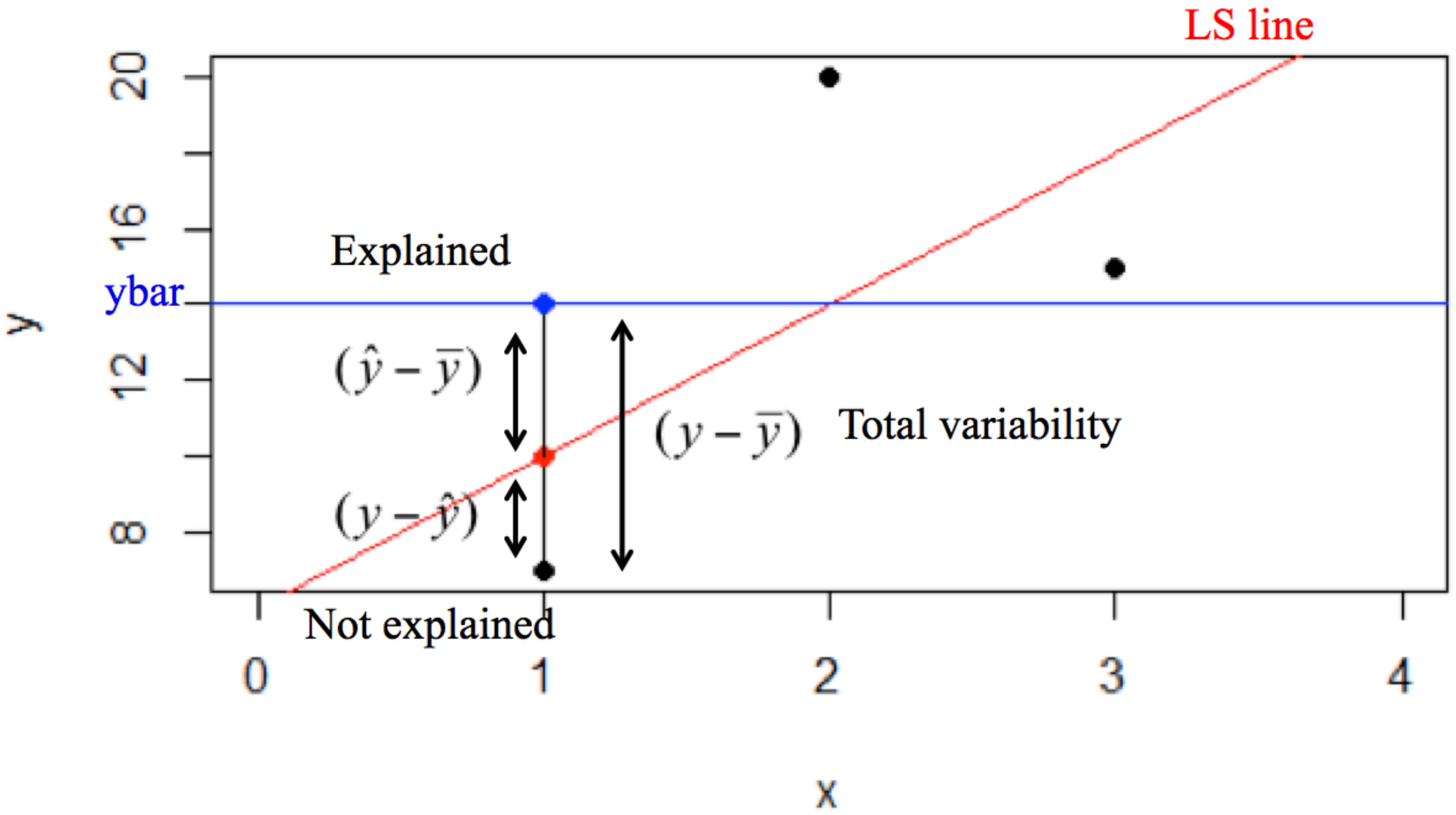
- The *coefficient of determination*, R^2, is the square of the Pearson correlation coefficient r (i.e., $R^2 = r^2$)
- R^2 represents the proportion of variability in y that is explained by its linear relationship with x
- Since $-1 \leq r \leq 1$, we have that $0 \leq R^2 \leq 1$
- Extremes:
  - If $R^2 = 1$, then all points lie on the regression line
  - If $R^2 = 0$, then there is no linear relationship between x and y

Coefficient of Determination

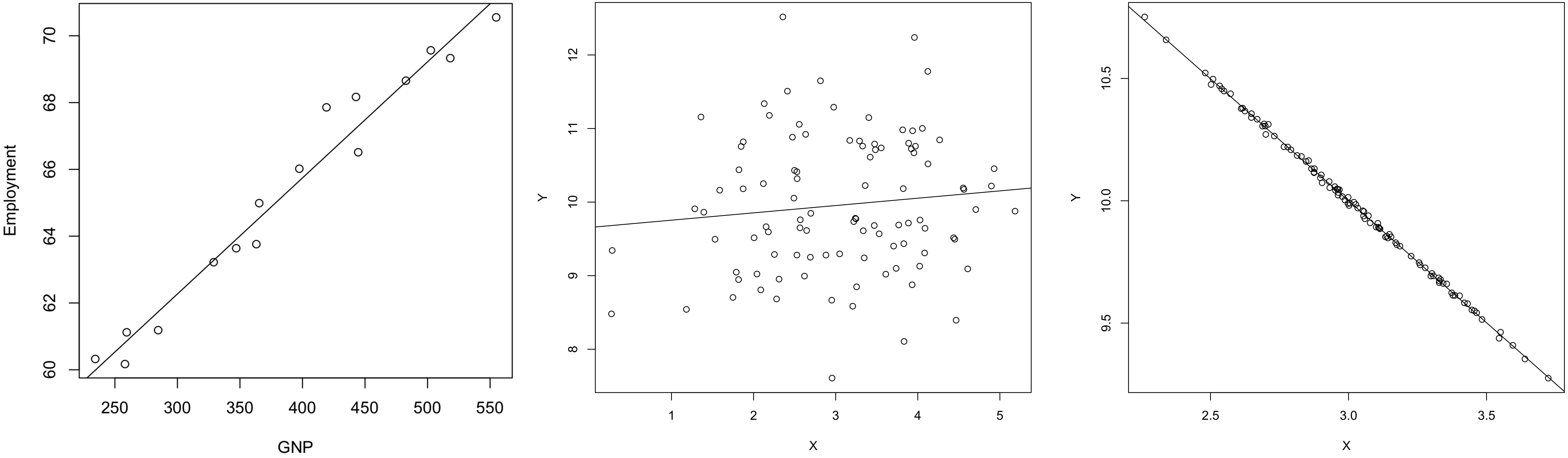
- R^2 has a few different equivalent representations
    - $R^2 = r^2$, where the correlation r is calculated between y and $\hat{y}$
    - $R^2 = \dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$ (the fraction of total squared error explained by $\hat{y}$)
    - $R^2 = 1 - \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \dfrac{s^2}{s_y^2}$, or $1 - \dfrac{SSE}{SSTo}$

    (SSE = SS explained, SSTo = SS total)

Coefficient of Determination



Coefficient of Determination

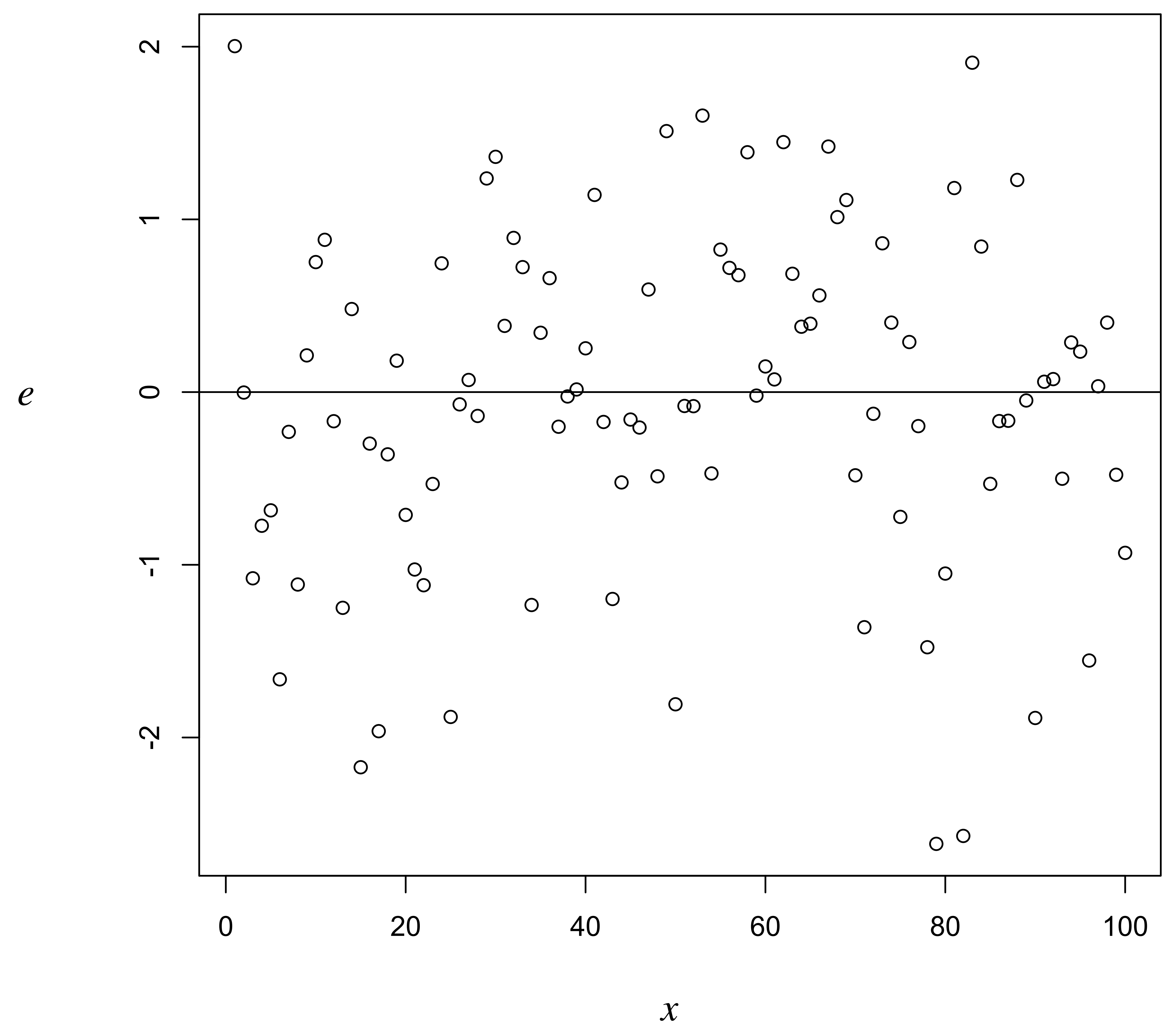


$R^2 = 0.96734$ $R^2 = 0.0121$ $R^2 = 0.9999$

Residual Plot

- Another way of evaluating model fit is through a *residual plot*
- Recall that Residuals = Actual – Predicted
- A residual plot is a scatterplot of the residuals over the fitted values, $\hat{y}_i$
- If an observed y_i is close to the fitted value $\hat{y}_i$, then the residual, $e_i = y_i - \hat{y}_i$, will be close to 0
- The estimated regression line is a good fit if we see random scatter in the residual plot around the line $x = 0$

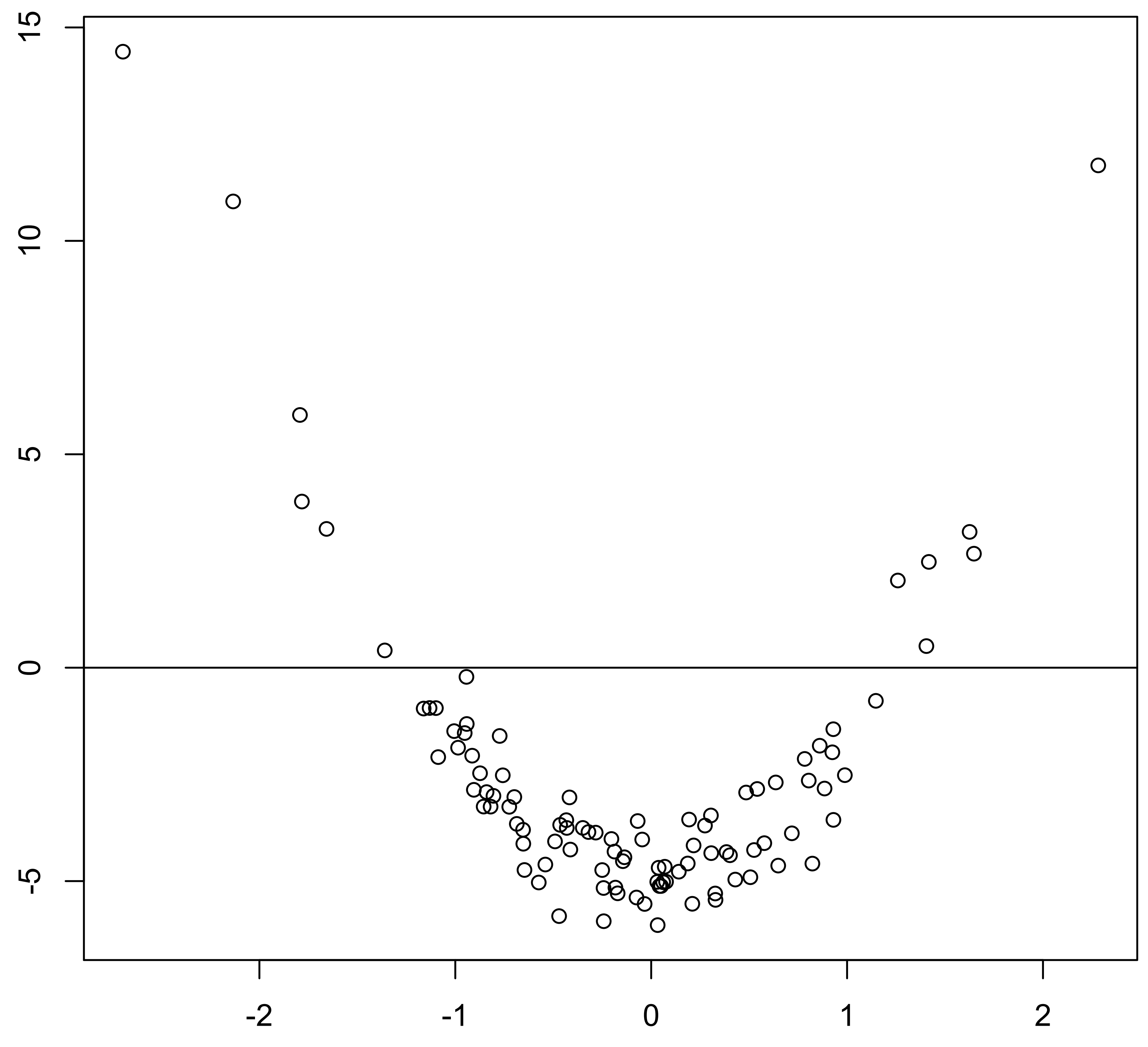
Residual Plot

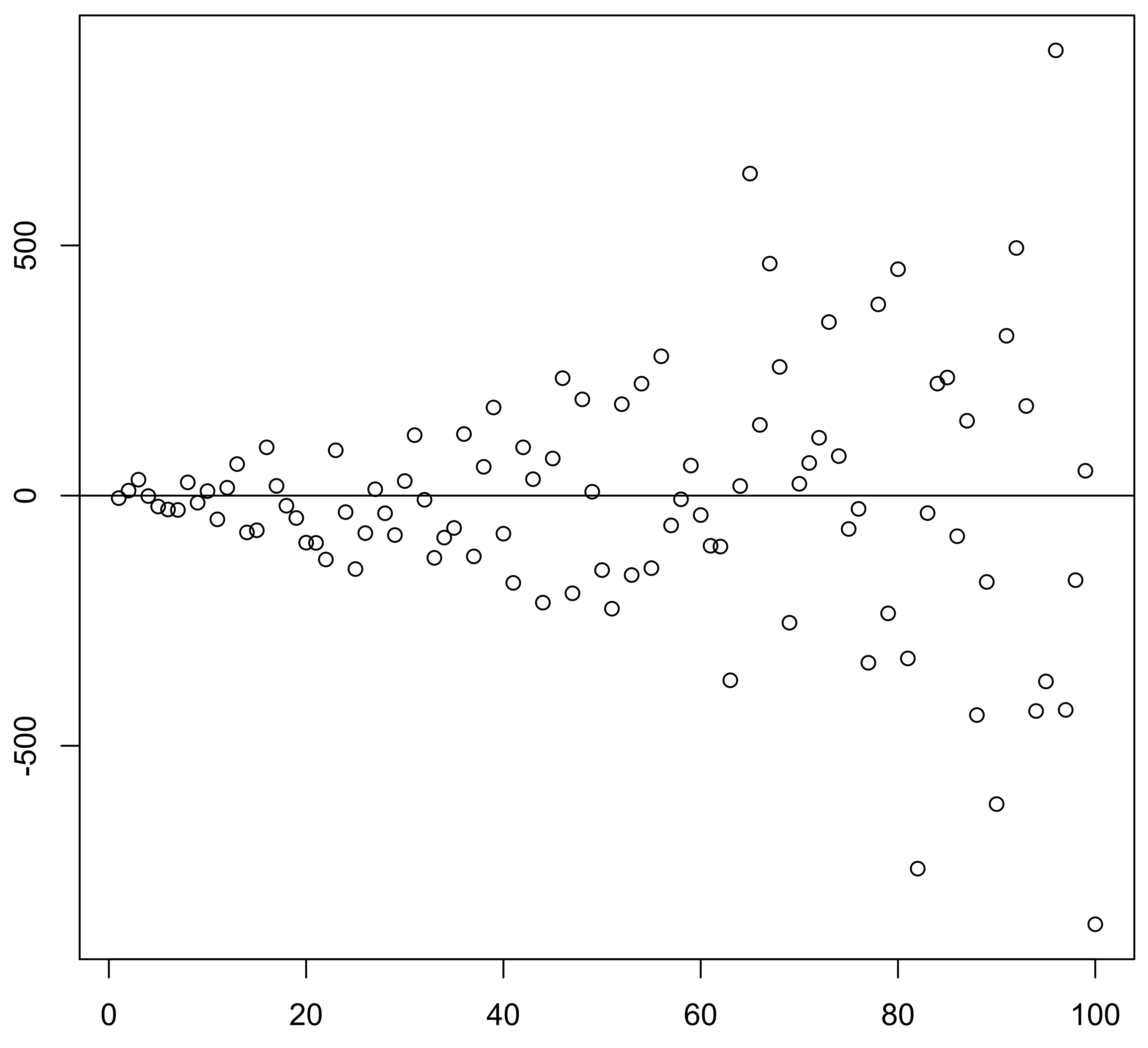


Residual Plot

- If we do not see random scatter but instead notice an obvious pattern, this indicates that our regression line is not too appropriate for modeling the data
- Some possible explanations:
  - Relationship is non-linear
  - Homoscedasticity assumption is not met, meaning that we do not have constant variance
- Can also use normal quantile-quantile (QQ) plots to assess the normality of the error terms

Residual Plot





Residual Plot

- If the residuals do not exhibit random scatter but instead appear to follow some trend, then the relationship between x and y is likely not linear
- A transformation of x or y (or both) may lead to a linear relationship
    - E.g., while x and y are not linearly related, perhaps $\ln(x)$ (natural log of x) and y may be linearly related

Multiple Linear Regression

- We've looked at how a single variable affects a response
  - E.g., how does age affect weight?
- What about multiple variables?
  - E.g., how do age, height, and eye color affect weight?
- We can extend previous methods to include multiple predictors

Multiple Linear Regression

- The regression model can now be written as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_q x_{qi} + \epsilon_i$ for $i = 1, \ldots, n$, $\epsilon \sim N(0, \sigma^2)$, and q predictor variables

- β_0 is the mean value of y when all predictors equal 0

- The slope β_j is the expected change in the mean value of y corresponding to a one-unit increase in x_j, *given that all other predictors are held constant*

Multiple Linear Regression

- Assumptions:
  - Given $x_1, \ldots, x_q$, the y's are independent
  - There is a linear relationship between y and $x_1, \ldots, x_q$ (i.e., $E(\epsilon) = 0$
  - The variance σ^2 is constant across all values of $x_1, \ldots, x_q$ (i.e., $Var(\epsilon) = \sigma^2$), known as homoscedasticity
  - For specified values of $x_1, \ldots, x_q$, y has a normal distribution
  - $x_1, \ldots, x_q$ are fixed, known quantities
- When these regression assumptions are met, the use of linear regression is appropriate for describing the relationship between y and $x_1, \ldots, x_q$

Multiple Linear Regression

- To fit the least squares regression line $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_q x_{qi}$, we again want to minimize the sum of squares of the residuals:

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_q x_{qi})^2$$

- The same form to calculate $\hat{\beta}_0$ and $\hat{\beta}_1, \ldots, \hat{\beta}_q$

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \ldots - \hat{\beta}_q \bar{x}_q$$

Multiple Linear Regression

- Consider a model for weight that depends on height and age
- Let $y =$ weight, $x_1 =$ height, and $x_2 =$ age
- This model assumes that both age and height linearly affect a person's weight
- We estimate the model parameters based on a sample of $n = 100$ subjects

Multiple Linear Regression

```
> lm1 <- lm(y~height+age)
> summary(lm1)

Call:
lm(formula = y ~ height + age)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5812 -0.6113  0.1729  0.6041  2.6114 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.91906    3.42299   4.943 3.22e-06 ***
height       1.99748    0.05384  37.102  < 2e-16 ***
age          0.08406    0.01091   7.706 1.13e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9734 on 97 degrees of freedom
Multiple R-squared:  0.9346,Adjusted R-squared:  0.9332 
F-statistic: 692.7 on 2 and 97 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

- The estimated regression line is $\hat{y} = 16.919 + 1.997x_1 + 0.084x_2$
- For a person who is 0 inches tall and of age 0, average weight is 16.919 pounds according to our model
- In this context, the y-intercept does not make sense and is extrapolating beyond the scope of our data
- Holding age constant, for each 1 inch increase in height, weight is expected to increase by 1.997 pounds
- Holding height constant, for each 1 year increase in age, weight is expected to increase by 0.084 pounds

Multiple Linear Regression: Inference

- We want to use the regression model to make inferences about the true population regression
- Let's first start with making inferences about a single parameter β_j at significance level α
- Hypotheses: $H_0 : \beta_j = \beta_j^*$ vs. $H_1 : \beta_j \neq \beta_j^*$ for some population value β_j^*
- Calculate $t = \dfrac{\hat{\beta}_j - \beta_j^*}{SE(\hat{\beta}_j)}$ and compare to a t-distribution with $n - q - 1$ degrees of freedom
- Calculate p-value: $p = \Pr(|T| \geq |t|) =$ `2*pt(-abs(t),df=n-q-1`
- If $p \leq \alpha$, reject the null hypothesis

Multiple Linear Regression: Inference

```
> lm1 <- lm(y~height+age)
> summary(lm1)

Call:
lm(formula = y ~ height + age)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5812 -0.6113  0.1729  0.6041  2.6114 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.91906    3.42299   4.943 3.22e-06 ***
height       1.99748    0.05384  37.102  < 2e-16 ***
age          0.08406    0.01091   7.706 1.13e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9734 on 97 degrees of freedom
Multiple R-squared:  0.9346,Adjusted R-squared:  0.9332 
F-statistic: 692.7 on 2 and 97 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression: Inference

- We can also create the following in the same manner as we did in the case of simple linear regression, but with degrees of freedom changing to $n - q - 1$ (recall that for simple linear regression, $q = 1$, so $n - q - 1 = n - 2$):
  - Confidence intervals for individual regression parameters
  - Confidence intervals for predicted mean values of y for fixed values of $x_1, \ldots, x_q$
  - Confidence intervals for a predicted individual y for fixed values of $x_1, \ldots, x_q$

Multiple Linear Regression: Inference

- Additionally, we can create an ANOVA table for multiple regression models

Source	SS	df	MS	F
Regression	SSR	q	$MSR = \frac{SSR}{q}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-q-1	$MSE = \frac{SSE}{n-q-1}$	
Total	SSTo	n-1		

- In this case, we can use the F statistic to test hypotheses about the values of all β_i's (not just one at a time)

Multiple Linear Regression: Inference

- In this case, the F statistic is used to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$ against the alternative hypothesis that at least one of these β_i values is nonzero
- This F statistic follows an F distribution with q and $n - q - 1$ degrees of freedom
- Calculate p-values using this F statistic (area in upper tail)

Multiple Linear Regression: Inference

- We've looked at inference for one variable (i.e., $H_0 : \beta_i = \beta_i^*$) and all variables together (i.e., $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$)
- What about for a subset of variables? Given two models, one of which is a submodel of the other, do the added predictors help give the larger model more predictive power, or are they extraneous?
    - Can also apply an F test here

Multiple Linear Regression: Inference

- Let the full model be as follows:
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \epsilon$
- For some $p < q$, the reduced model is
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$
- The reduced model is obtained by removing the final $q - p$ predictors from the full model
- In this case, we say that the reduced model is *nested* in the full model

Multiple Linear Regression: Inference

- The goal is to determine if the reduced model is sufficient or if we gain predictive ability by adding in the final $q - p$ predictors

- In other words, we want to test the following:

$$H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_q = 0$$

$$H_1 : \text{at least one of these equalities does not hold}$$

- Our test statistic is as follows:

$$F = \frac{(SSE_p - SSE_q)/(q - p)}{SSE_q/(n - q - 1)}$$

- F follows an F distribution with $q - p$ and $n - q - 1$ degrees of freedom

Multiple Linear Regression: Model Evaluation

- We can evaluate the fit of the regression model through the adjusted R^2 and residual plots
- Residual plots can be created as before and used for judging whether the model is appropriate, as in the case of single linear regression
- If the residual plots do not display random scatter, this indicates that y is not linearly related to $x_1, \ldots, x_q$

Multiple Linear Regression: Adjusted R^2

- However, for multiple linear regression, we use the adjusted R^2 instead of R^2
- Intuition: adjusted R^2 penalizes the use of more explanatory variables
    - By adding more predictors to the regression model, we cannot make our model fit worse
    - Adjusted R^2 only increases when an added variable improves our ability to predict the response (only rewards useful explanatory variables)
    - Does not have the same interpretation as R^2
- Formula: Adjusted $R^2 = \overline{R}^2 = 1 - \dfrac{SSE/df_E}{SST/df_T} = 1 - (1 - R^2) \cdot \dfrac{n-1}{n-q}$

Multiple Linear Regression: Indicator Variables

- Some predictor variables may be categorical instead of continuous, which we have not considered so far
  - E.g., include sex as a predictor of weight
- In a regression model, predictor variables must take numerical values, so we assign arbitrary integer values to categories
  - E.g., male = 1, female = 0
- Since the values of these variables do not have any direct meaning, we refer to these variables as *indicator* or *dummy* variables

Multiple Linear Regression: Indicator Variables

- Let's suppose we now use height, age, and sex to linearly predict weight
- Let x_3 = sex (for this example, assume binary)
- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$
- Since $x_{3i} = 0$ if subject i is female and $x_{3i} = 1$ if subject i is male, $\hat{\beta}_3$ is the estimated difference in mean weight for males compared to females
- Always compare to the reference group (i.e., $x_j = 0$)
- $\hat{\beta}_0$ is then the estimated weight for a woman of height 0 and age 0

Multiple Linear Regression: Indicator Variables

```
> lm1 <- lm(y~height+age+sex)
> summary(lm1)

Call:
lm(formula = y ~ height + age + sex)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5319 -0.5742  0.1548  0.6494  2.1226 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.35557    3.47273   0.678    0.499    
height       2.22447    0.05484  40.562  < 2e-16 ***
age          0.09777    0.01101   8.876 3.86e-14 ***
sex1         1.90646    0.19978   9.543 1.43e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9828 on 96 degrees of freedom
Multiple R-squared:  0.9515,	Adjusted R-squared:   0.95 
F-statistic: 627.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression: Indicator Variables

- Indicator variables only change the y-intercept of the regression line
- For a female, the regression line is

$$\begin{aligned}\hat{y} &= 2.356 + 2.224x_1 + 0.098x_2 + 1.906(0) \\ &= 2.356 + 2.224x_1 + 0.098x_2\end{aligned}$$

- For a male, the regression line is

$$\begin{aligned}\hat{y} &= 2.356 + 2.224x_1 + 0.098x_2 + 1.906(1) \\ &= 4.262 + 2.224x_1 + 0.098x_2\end{aligned}$$

- Thus, men, on average, have higher weights than women

Multiple Linear Regression: Interactions

- This is assuming that the indicator variable (in this case, sex) doesn't interact with any other explanatory variables
- However, sometimes it is beneficial to allow certain variables to depend on an indicator random variable
  - For instance, maybe it is reasonable to allow age to affect weight differently for men and women
- In this case, the slope of the regression line would be different for men and women, as well as the y-intercept
- In general, one predictor variable may have a different effect on the predicted response y depending on the value of a second predictor variable
- Allow for an *interaction term* by multiplying together the outcomes of the two predictors

Multiple Linear Regression: Interactions

```
> lm1 <- lm(y~height+age+sex+sex*age)
> summary(lm1)

Call:
lm(formula = y ~ height + age + sex + sex * age)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9771 -0.6721 -0.0454  0.8603  3.2796 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.463101   3.124635   1.748   0.0836 .  
height      2.178340   0.049102  44.364  < 2e-16 ***
age         0.089854   0.019372   4.638 1.12e-05 ***
sex1        1.428663   1.169072   1.222   0.2247    
age:sex1    0.008914   0.028999   0.307   0.7592    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.115 on 95 degrees of freedom
Multiple R-squared:  0.9576,    Adjusted R-squared:  0.9558 
F-statistic: 536.6 on 4 and 95 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression: Interactions

- Now, we have $\hat{y} = 5.463 + 2.178x_1 + 0.090x_2 + 1.429x_3 - 0.009x_2x_3$
- For a female, the regression line is

$$\begin{aligned}\hat{y} &= 5.463 + 2.178x_1 + 0.090x_2 + 1.429(0) - 0.009x_2(0) \\ &= 5.463 + 2.178x_1 + 0.090x_2\end{aligned}$$

- For a male, the regression line is

$$\begin{aligned}\hat{y} &= 5.463 + 2.178x_1 + 0.090x_2 + 1.429(1) - 0.009x_2(1) \\ &= 6.892 + 2.178x_1 + 0.081x_2\end{aligned}$$

- In this case, the interaction term is not very significant, indicating that we do not need to separately model age's effect for men and women

Multiple Linear Regression: Collinearity

- For any model with multiple variables, it is important to check for collinearity
- Two variables may be highly correlated and thus both should not be included in the model
- Standard errors for parameter estimates typically become large when collinearity is present
- Calculate the correlation between all predictor variables
- If two variables are highly correlated, you should consider removing the one that changes the model fit the least