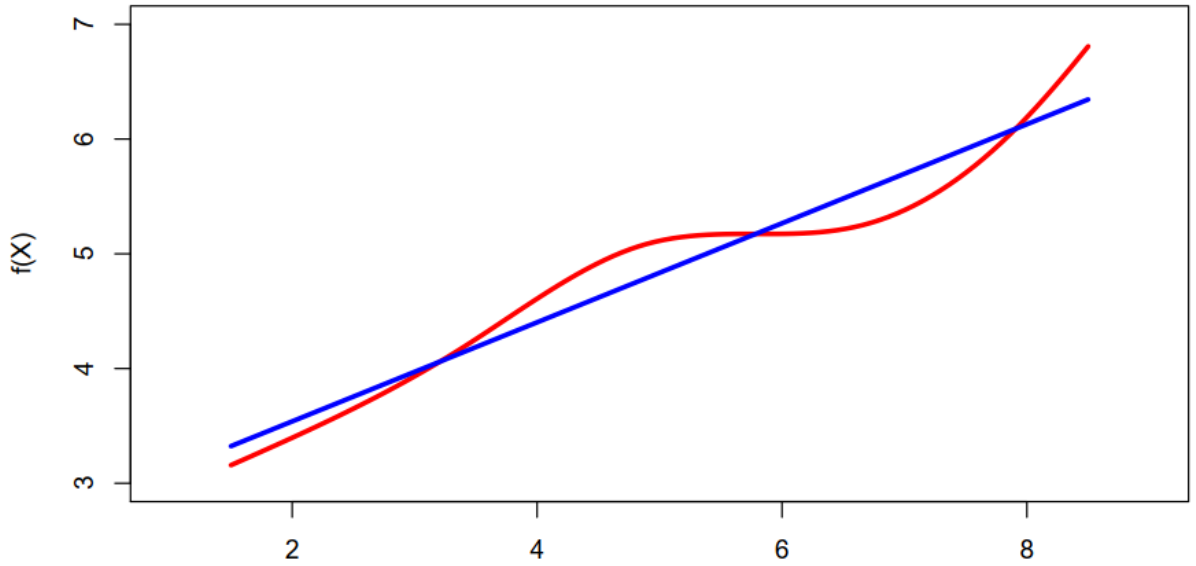
Regression

Szu-Chi Chung

Department of Applied Mathematics, National Sun Yat-sen University

Linear regression

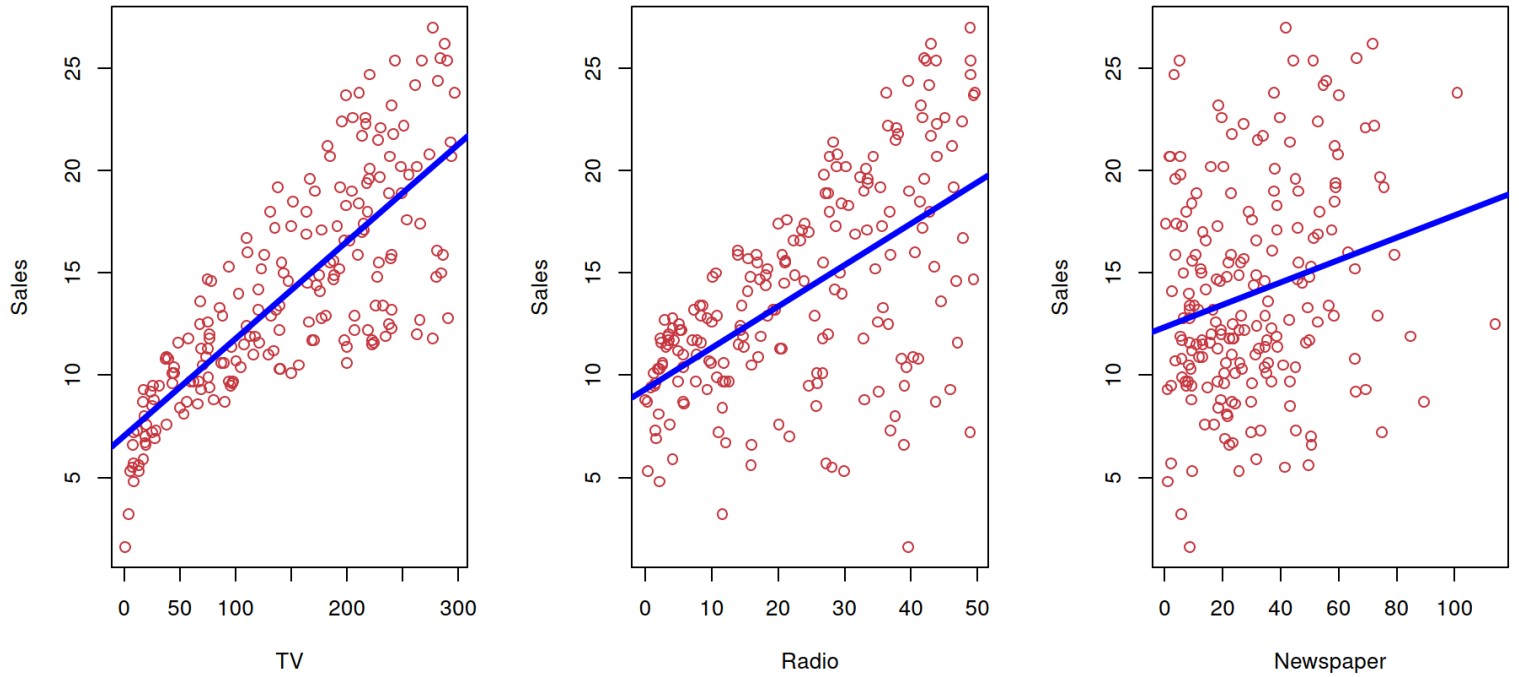
- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear
- True regression functions are never linear!



- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically

Linear regression for the advertising data

- Consider the advertising data, questions we might ask:
  - Is there a *relationship* between advertising budget and sales?
  - How *strong* is the relationship between advertising budget and sales?
  - Which media contribute to sales?
  - *How accurately* can we predict future sales?
  - Is the relationship linear?
  - Is there *synergy* among the advertising media?



Simple linear regression using a single predictor X

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as *coefficients* or *parameters*, and ϵ is the error term which is assumed to be i.i.d. that follows the normal distribution. (LINE)

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2), \sigma^2 = Var(\epsilon)$$

- Given some estimates β_0 and β_1 for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value

Estimation of the parameters by least squares

- There are many ways of measuring closeness. We use *least squares* here
  - Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
  - We define *the residual sum of squares (RSS)* as

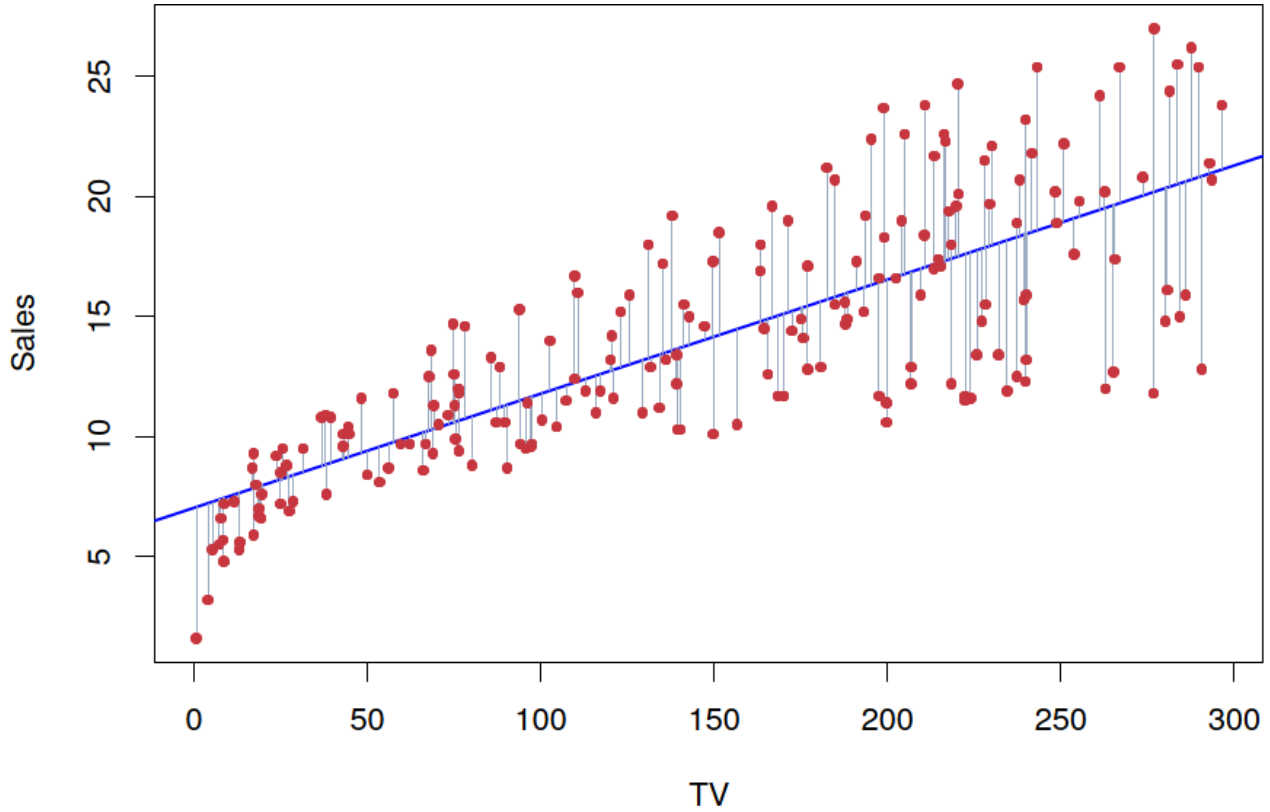
$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$
$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

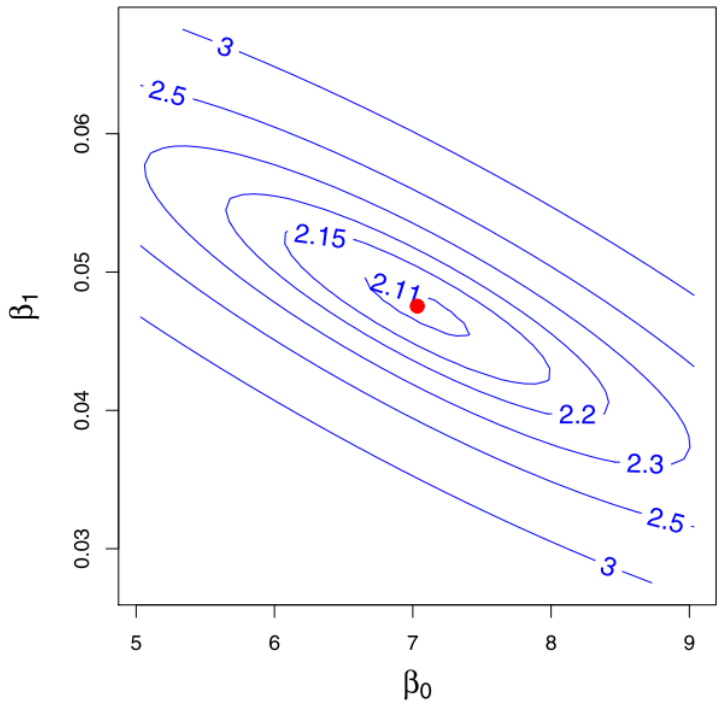
- The least squares approach (Maximum likelihood) chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

  - where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample mean
  - Scale does not affect the estimation of $\hat{y}$

Example: Advertising data

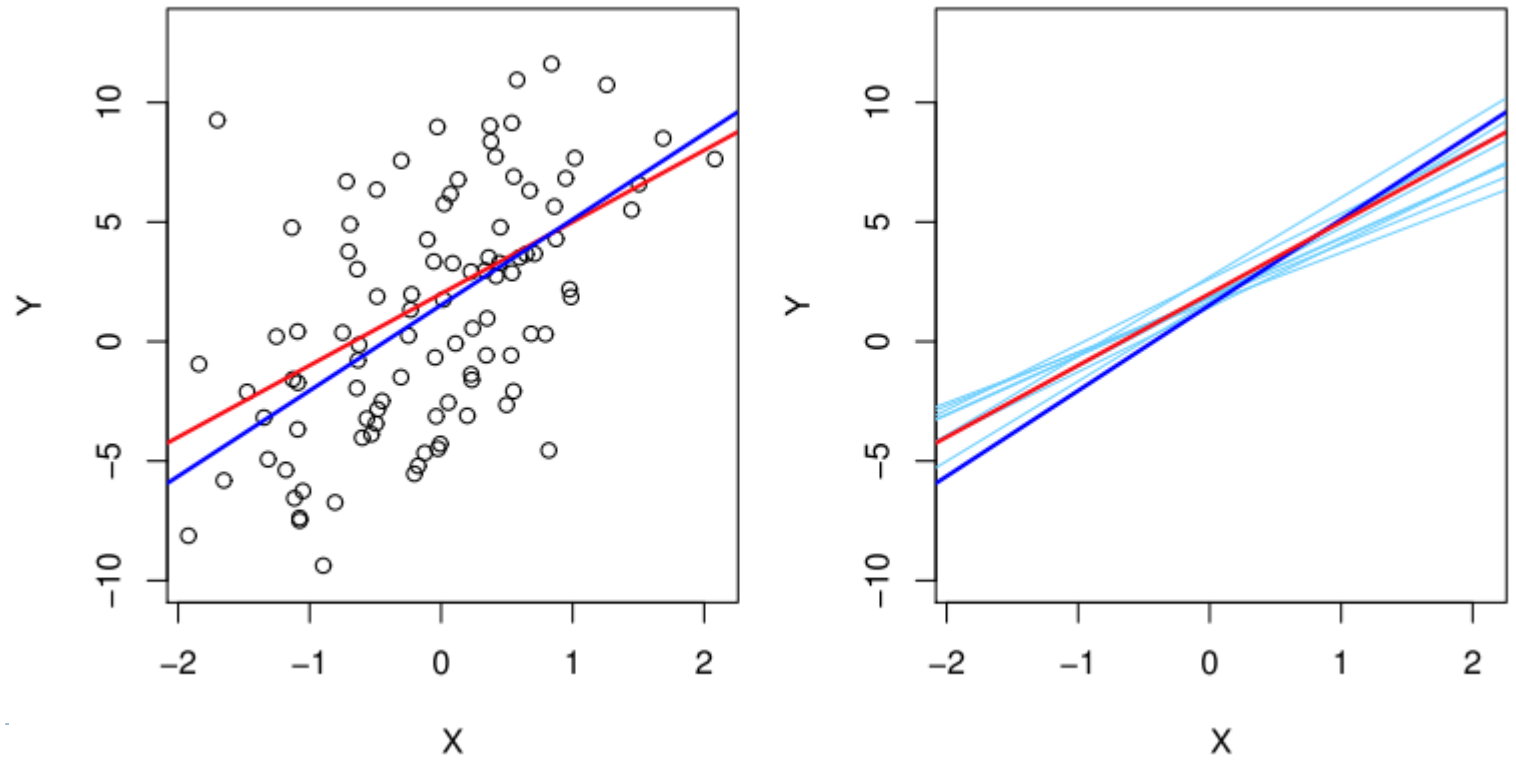




- The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is deficient in some part of the plot

(1) Assessing the Accuracy of the Coefficient Estimates

- $Y = 2 + 3X + \epsilon$
  - Red line indicates the *population regression line*
- For example, if we want to estimate the population mean and the *standard error*
  - $\hat{\mu} = \frac{1}{n}\sum x_i$
  - $Var(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$
  - σ is standard deviation of x_i's



Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

  - where $\sigma^2 = Var(\epsilon), \hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (RSE)^2$

- These standard errors can be used to compute *confidence intervals*
  - A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$$

Confidence intervals — continued

- That is, there is approximately a 95% chance that the interval

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

- For the advertising data, the 95% confidence interval for β_1 is [0.042, 0.053] and for β_0 is [6.130, 7.935]
  - In the absence of advertising, sales will on average fall somewhere between 6,130 and 7,935 units
  - For each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units

Hypothesis testing

- Standard errors can also be used to perform *hypothesis tests* on the coefficients

- The most common hypothesis test involves testing the *null hypothesis* of

H_0 : There is no relationship between X and Y

versus the alternative hypothesis

H_a : There is some relationship between X and Y

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y

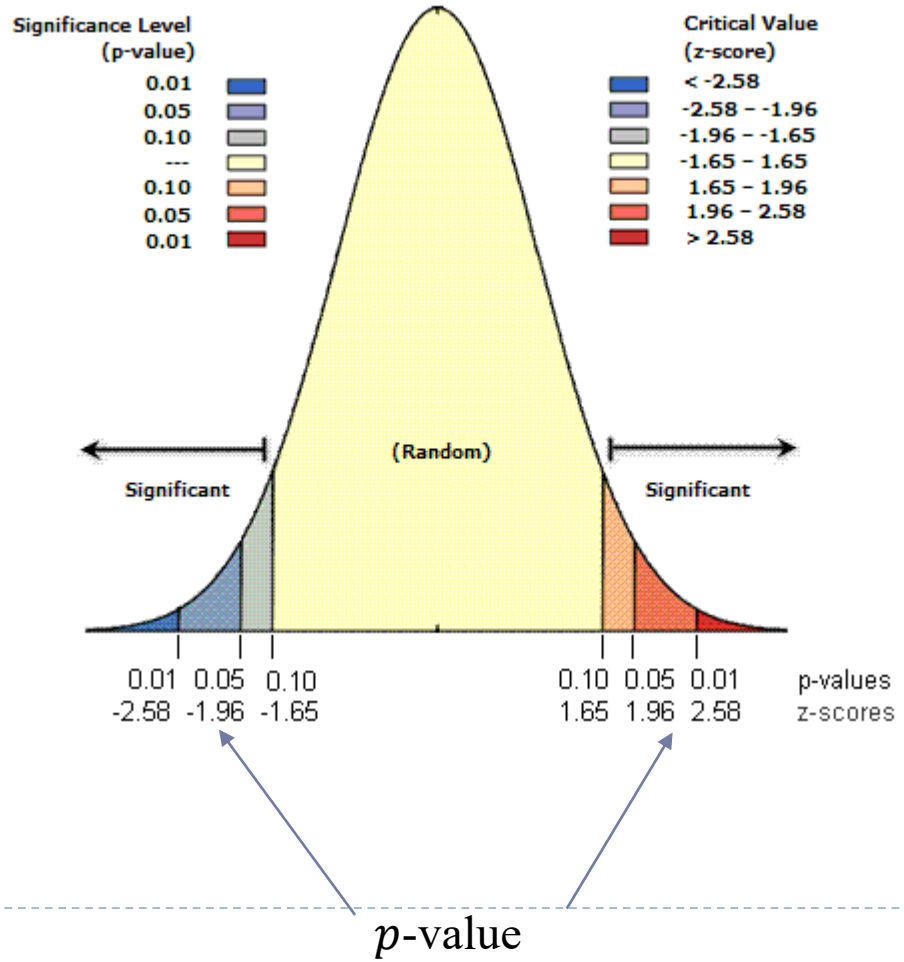
Hypothesis testing — continued

- To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

  - This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the *p-value*

https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/what-is-a-z-score-what-is-a-p-value.htm



(2) Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error* (RSE) (smaller is better)

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \hat{\sigma}$$

  where the residual sum-of-squares is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- *R-squared* or fraction of variance explained is (larger is better)

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the total sum of squares

- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y (Exercise 7):

$$r = Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{x})^2}}$$

Advertising data results

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

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

Extension of simple linear regression

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Multiple Linear Regression

- Instead of fitting three simple linear regression model for each predictor
- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, *holding all other predictors fixed*. In the advertising example, the model becomes

$$Sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

Estimation and Prediction for Multiple Regression

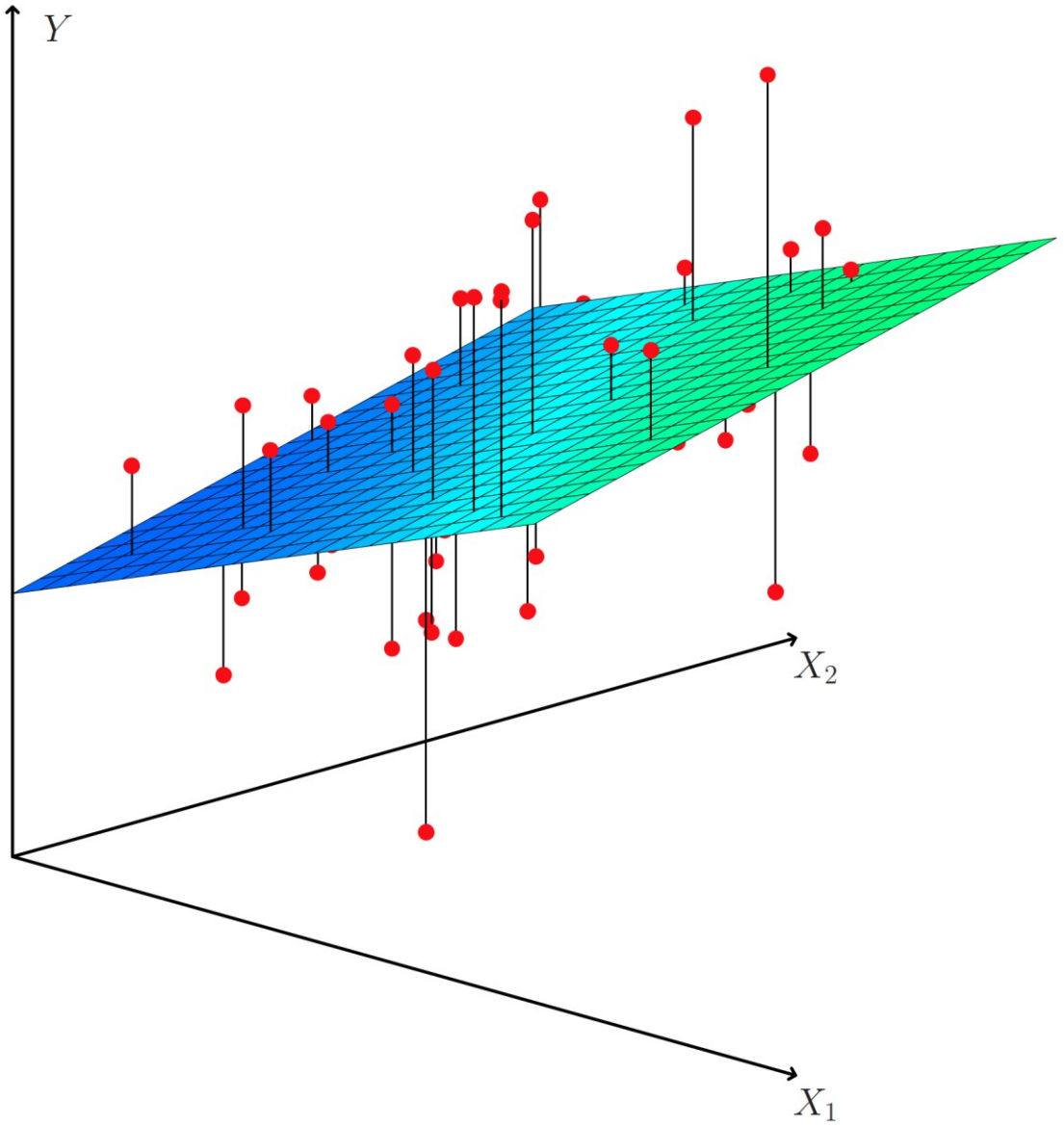
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- We estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ as the value that minimize the sum of squared residuals

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

- This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates


Y
X2
X1

Results for advertising data

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

Correlation matrix

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Some important questions

1. Is at least one of the predictors $X_1, X_2, \ldots X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is at least one predictor useful?

- For the first question, we can use the F-statistic

$$H_0 : \beta_1 = \beta_1 = \cdots = \beta_p = 0 \text{ versus } H_a : \text{at least one } \beta_j \neq 0$$

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

  - Note if linear model assumption is hold, $E\left\{\frac{RSS}{n-p-1}\right\} = \sigma^2$ and if H_0 hold, $E\{(TSS -$

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Is at least one predictor useful?

- To examine a particular set of q variables are zeros or not

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

$$\text{we use } F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

- We fit a second model that uses all the variables except those last q to get RSS_0
- In the previous table, $q = 1$
  - It seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response. Why should we care about the overall F-statistics?

Deciding on the important variables

- If $p > n$, we cannot fit the multiple linear regression model using least squares and therefore can not use F-statistics!
  - On the other hand, if we conclude on the basis of that p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that *balances training error with model size*
  - However we often can't examine all possible models, since they are 2^p of them; for example, when $p = 40$, there are over a billion models!

Deciding on the important variables - Forward selection

1. Begin with the null model — a model that contains an intercept but no predictors
2. Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS
3. Add to that model the variable that results in the lowest RSS amongst all two-variable models
4. Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold

Deciding on the important variables - Backward selection

1. Start with all variables in the model
2. Remove the variable with the largest p-value — that is, the variable that is the least statistically significant
3. The new $(p-1)$ −variable model is fit, and the variable with the largest p −value is removed
4. Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold

Deciding on the important variables - Mixed selection

1. Begin with the null model, and as with forward selection, we add the variable that provides the best fit
2. If at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model
3. Continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model

Deciding on the important variables - Mixed selection

- Mixed selection can remedy the following situations
  - Backward selection cannot be used if $p > n$, while forward selection can always be used
  - Forward selection is a greedy approach and might include variables early that later become redundant
- Later we discuss more systematic criteria for choosing an "optimal number" of members in the path of models produced by forward stepwise selection
- These include Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 and Cross-validation (CV)

Model fit and prediction

- It can be shown that $R^2 = Cor(Y, \hat{Y})$ in this case
  - R^2 will always increase when more variables are added to the model
- We compute the Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-p-1}RSS} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \hat{\sigma}$$

- How close between $\hat{Y}$ and $f(X)$ can be quantified by the *confidence interval*
  - Note that even if we knew $f(X)$—that is, even if we knew the true values for $\beta_0, \beta_1, \ldots, \beta_p$—the response value cannot be predicted perfectly because of the random error ϵ (irreducible error)
  - For the new prediction, how much will Y vary from $\hat{Y}$? We use *prediction intervals* to answer this question

Confidence interval versus prediction interval

- Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error)

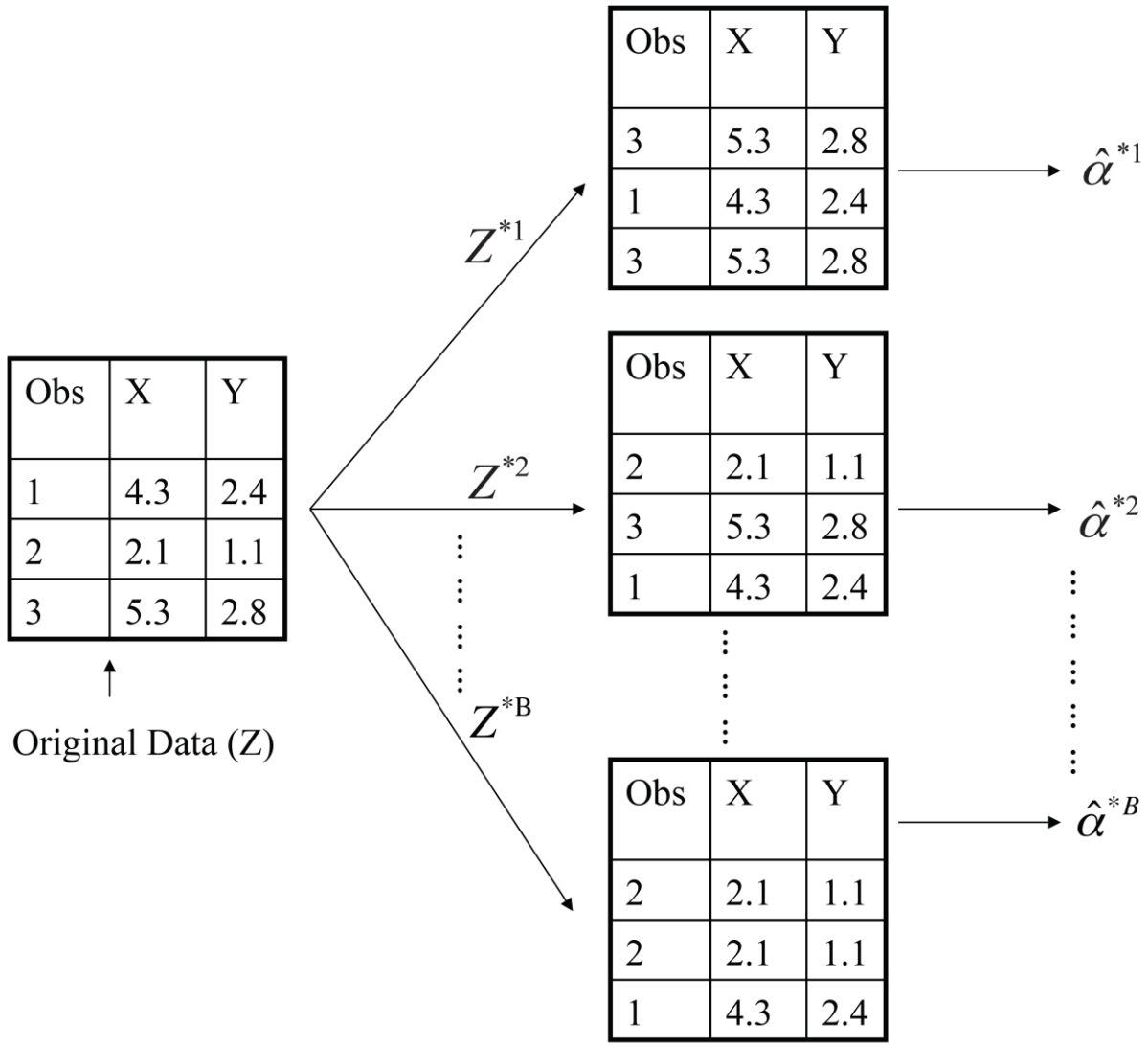
$$C.I. = \hat{y}_i \pm t_{\alpha/2}\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$P.I. = \hat{y}_i \pm t_{\alpha/2}\hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

  - For multiple linear regression see here

Bootstrap for confidence interval

- A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations
  - Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α



Bootstrap for confidence interval

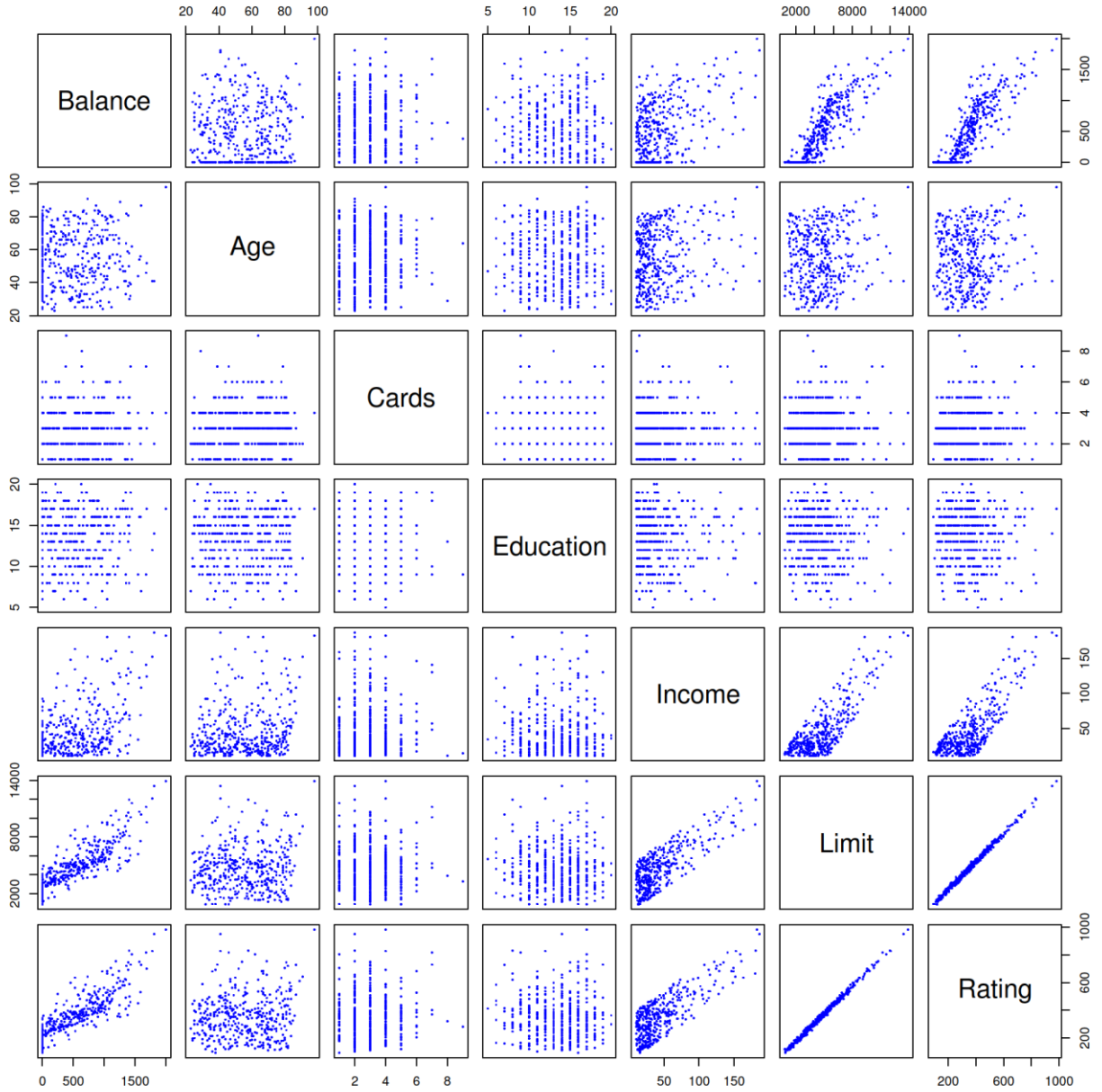
https://www.statsmodels.org/stable/examples/notebooks/generated/lowess.html#Confidence-interval

1. Generate n "bootstrap sample" data points x_i^*, y_i^*
2. Fit linear regression using x_i^*, y_i^*
3. Evaluate the regression line on fix x-grid
4. Repeat step 1-3 for B times and collect the values in step 3.
5. For each point in the x-grid, calculate the confidence interval using collected value

```python
def lowess_with_confidence_bounds(
    x, y, eval_x, N=200, conf_interval=0.95, lowess_kw=None
):
    """
    Perform Lowess regression and determine a confidence interval by bootstrap resampling
    """
    # Lowess smoothing
    smoothed = sm.nonparametric.lowess(exog=x, endog=y, xvals=eval_x, **lowess_kw)

    # Perform bootstrap resamplings of the data
    # and  evaluate the smoothing at a fixed set of points
    smoothed_values = np.empty((N, len(eval_x)))
    for i in range(N):
        sample = np.random.choice(len(x), len(x), replace=True)
        sampled_x = x[sample]
        sampled_y = y[sample]

        smoothed_values[i] = sm.nonparametric.lowess(
            exog=sampled_x, endog=sampled_y, xvals=eval_x, **lowess_kw
        )

    # Get the confidence interval
    sorted_values = np.sort(smoothed_values, axis=0)
    bound = int(N * (1 - conf_interval) / 2)
    bottom = sorted_values[bound - 1]
    top = sorted_values[-bound]

    return smoothed, bottom, top


# Compute the 95% confidence interval
eval_x = np.linspace(0, 4 * np.pi, 31)
smoothed, bottom, top = lowess_with_confidence_bounds(
    x, y, eval_x, lowess_kw={"frac": 0.1}
)
```

Other Considerations in the Regression Model

- Qualitative Predictors
  - Some predictors are not quantitative but are qualitative, taking a discrete set of values
  - These are also called *categorical* predictors or *factor* variables
  - See for example the scatterplot matrix of the credit card data in the next slide
  - In addition to the 7 quantitative variables shown, there are four qualitative variables: own (house ownership), student (student status), status (marital status), and region (East, West or South)


Balance
Age
Cards
Education
Income
Limit
Rating

Qualitative Predictors — continued

- Example: investigate differences in credit card balance between a person who has a house or not, ignoring the other variables. We create a *dummy variable*

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house} \end{cases}$$

- Resulting model:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does noy own a house} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`own[Yes]`	19.73	46.05	0.429	0.6690

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the region variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West} \end{cases}$$

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East} \end{cases}$$

Qualitative predictors with more than two levels

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — East in this example — is known as the *baseline*

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`region[South]`	−18.69	65.02	−0.287	0.7740
`region[West]`	−12.50	56.68	−0.221	0.8260

Extensions of the Linear Model

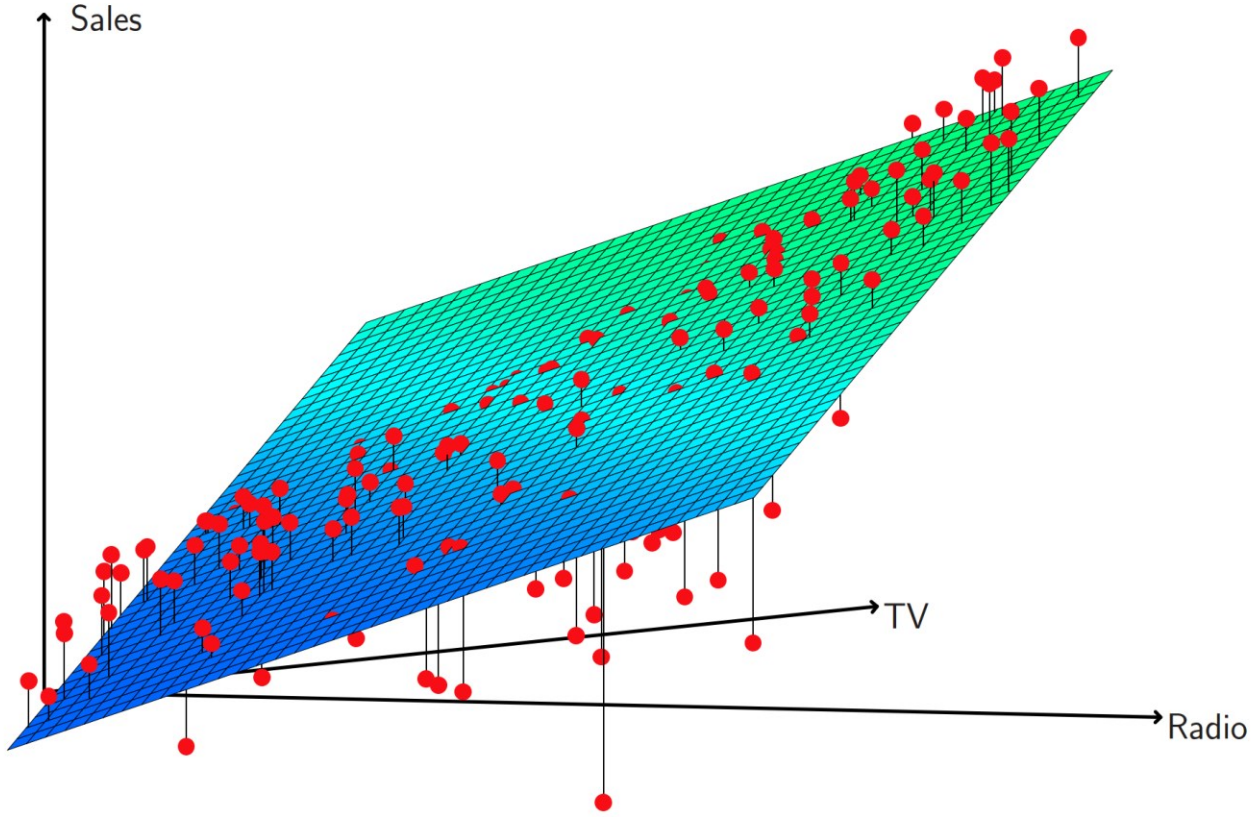
- Removing the additive assumption: interactions and nonlinearity
- Interactions:
  - In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media
  - For example, the linear model

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper$$

  states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio
- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases

Extensions of the Linear Model

- Given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio
  - In marketing, this is known as a *synergy effect*, and in statistics it is referred to as an *interaction effect*
  - The positive residuals (those visible above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals, tend to lie away from this line



Modelling interactions — Advertising data

- Model takes the form

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \epsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + \epsilon$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

  - The results in this table suggests that interactions are important
  - The p-value for the interaction term $radio \times TV$ is extremely low, indicating that there is strong evidence for H_a: $\beta_3 \neq 0$
  - The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term

Modelling interactions — Advertising data

- This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term
  - The coefficient estimates in the table suggest that an increase in TV advertising of \$1, 000 is associated with increased sales of $(\beta_1 + \beta_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$ units
  - An increase in radio advertising of \$1, 000 will be associated with an increase in sales of $(\beta_2 + \beta_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$ units
- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not
  - The *hierarchy principle*: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant. Otherwise the interpretation may change

Interactions between qualitative and quantitative variables

- Consider the Credit data set, and suppose that we wish to predict balance using income (quantitative) and student (qualitative)
- Without an interaction term, the model takes the form

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

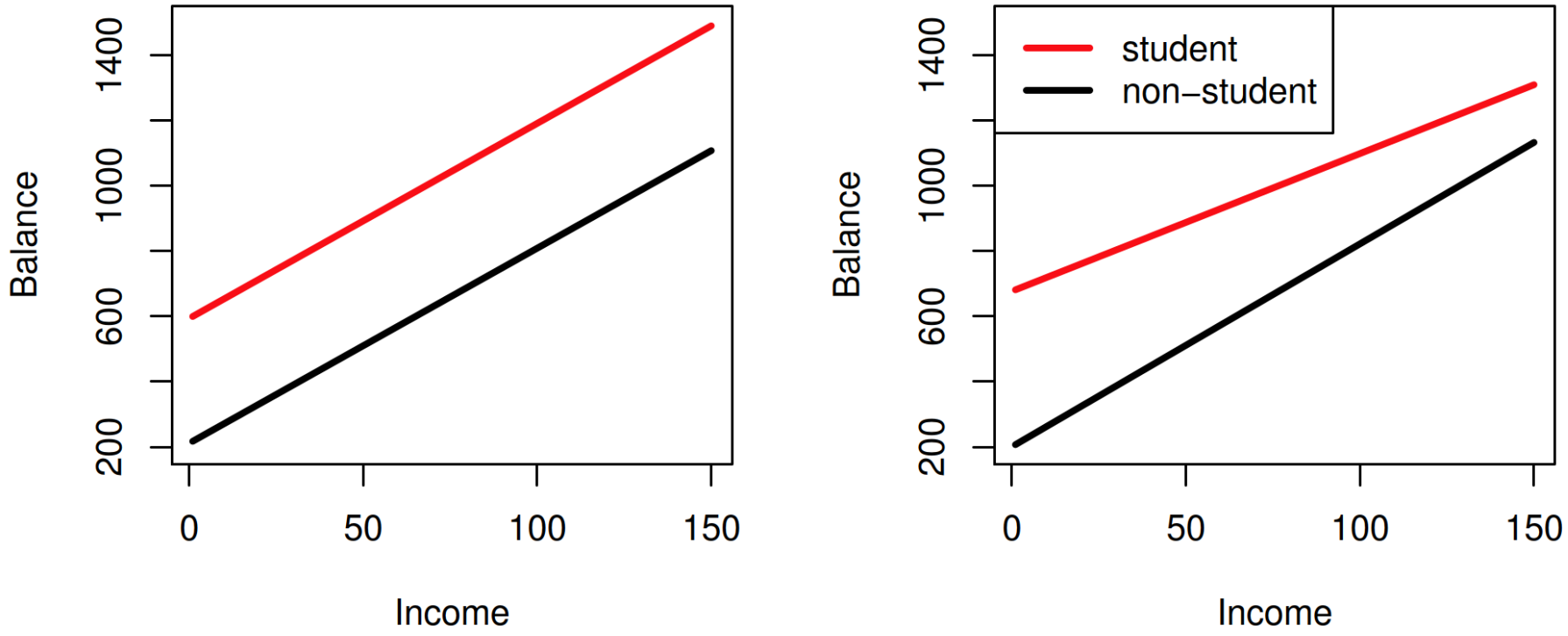
$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases}$$

- With interactions, it takes the form

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 + \beta_3 \times income_i & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income_i & \text{if } i\text{th person is a student} \\ \beta_0 + \beta_1 \times income_i & \text{if } i\text{th person is not a student} \end{cases}$$

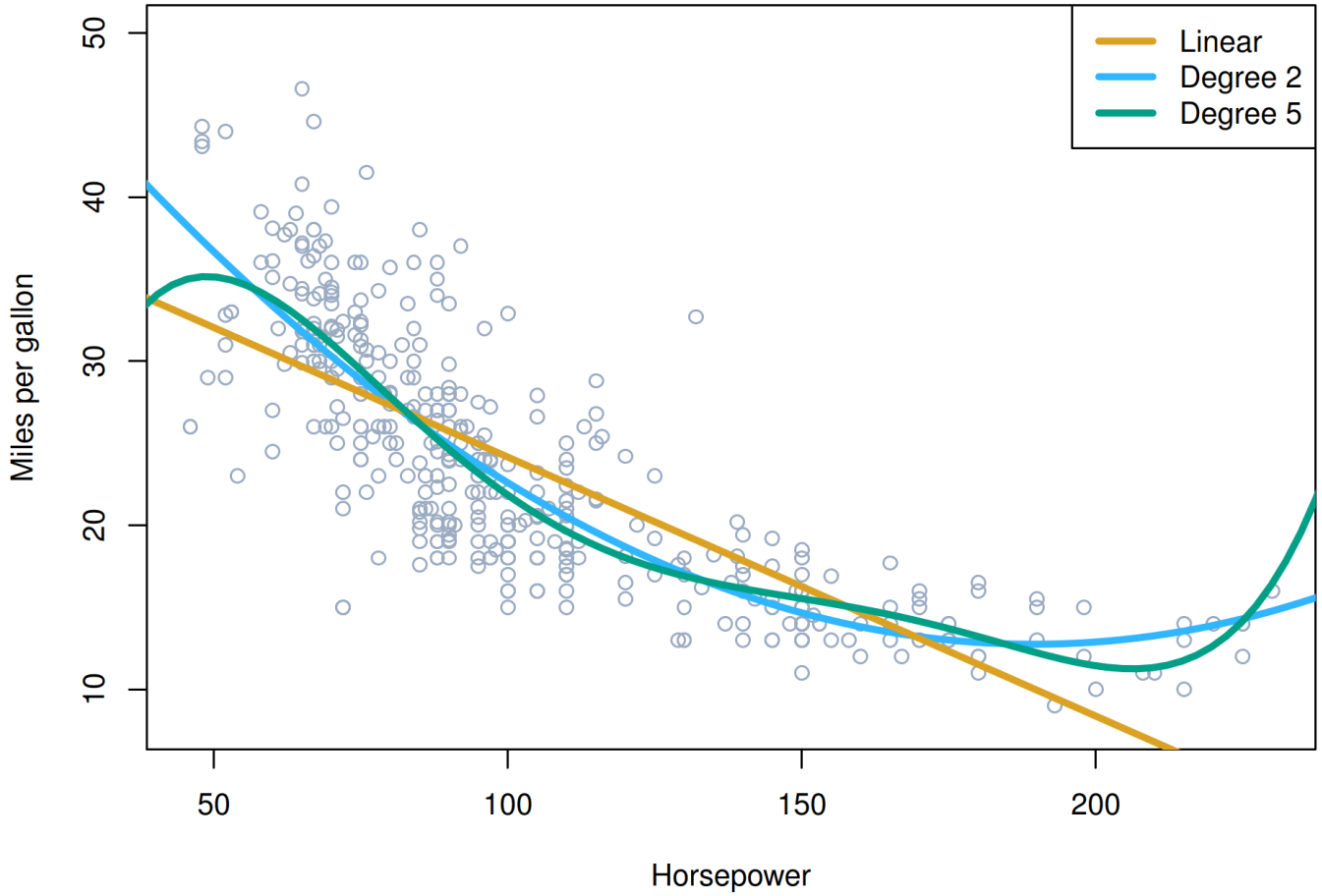
Interactions between qualitative and quantitative variables



Credit data; Left: no interaction between income and student
Right: with an interaction term between income and student

Non-linear effects of predictors

- *Polynomial regression* on Auto data



Non-linear effects of predictors

- The figure suggests that

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon$$

may provide a better fit

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
`horsepower`^{2}	0.0012	0.0001	10.1	< 0.0001

- The R^2 of the quadratic fit is 0.688, compared to 0.606 for the linear fit, and the p-value for the quadratic term is highly significant
- If including horsepower^2 led to such a big improvement in the model, why not include horsepower^3, horsepower^4, or even horsepower^5?

Potential Problems

- When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:
  1. Non-linearity of the response-predictor relationships
  2. Correlation of error terms
  3. Non-constant variance of error terms
  4. Outliers
  5. High-leverage points
  6. Collinearity

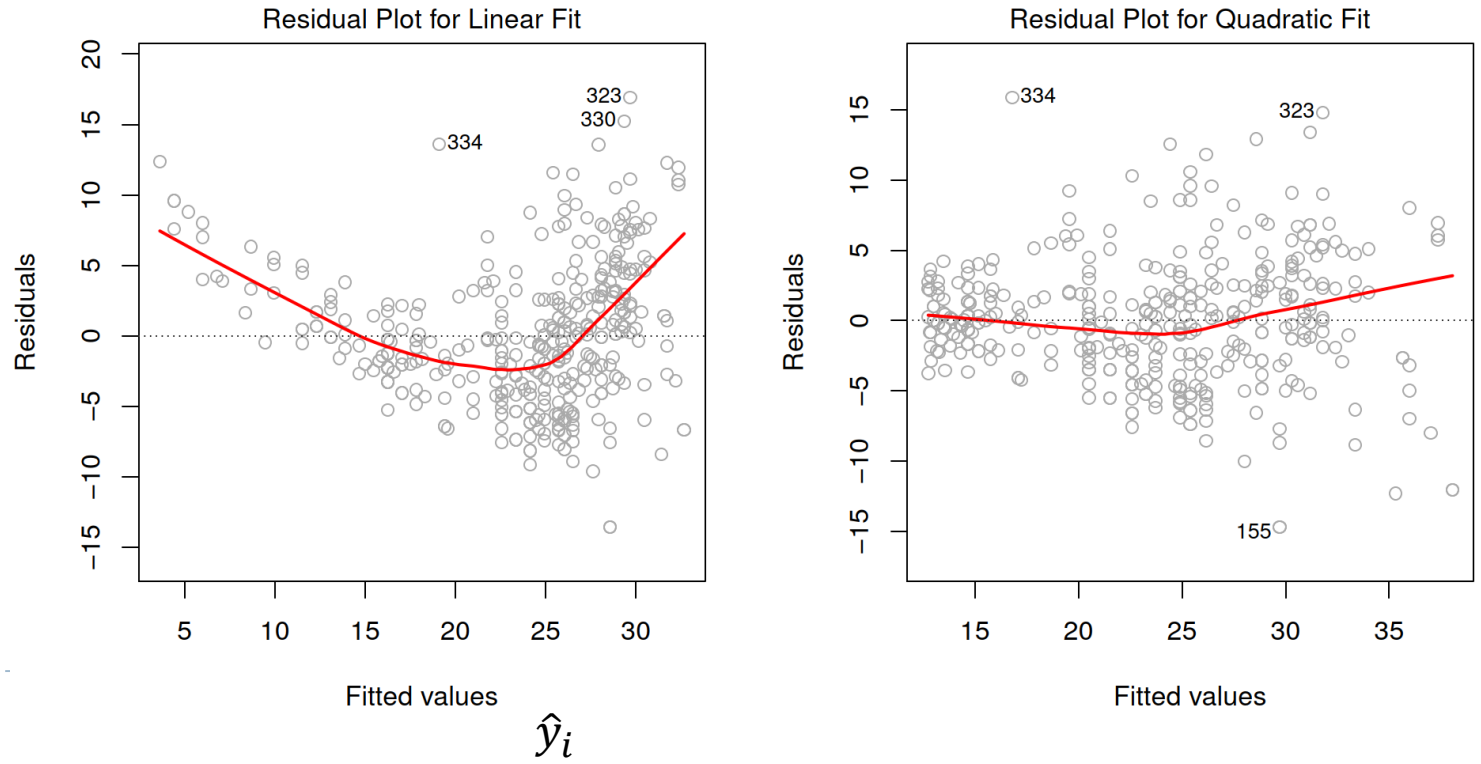
In practice, identifying and overcoming these problems is as *much an art as a science*

1. Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response
  - If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect
  - In addition, the prediction accuracy of the model can be significantly reduced
- *Residual plots* are a useful graphical tool for identifying non-linearity
  - Ideally, the residual plot will show no discernible pattern. The presence of a pattern may indicate a problem with some aspects of the linear model
  - The left panel of Figure 3.9 displays a residual plot from the linear regression of mpg onto horsepower on the Auto data set that was illustrated in Figure 3.8

- The red line is a smooth fit to the residuals, which is displayed in order to make it easier to identify any trends
- The residuals exhibit a clear U-shape, which provides a strong indication of non-linearity in the data
- In contrast, the right-hand panel of Figure 3.9 displays the residual plot that results from the model (3.36), which contains a quadratic term
- There appears to be little pattern in the residuals, suggesting that the quadratic term improves the fit to the data

$e_i = y_i - \hat{y}_i$



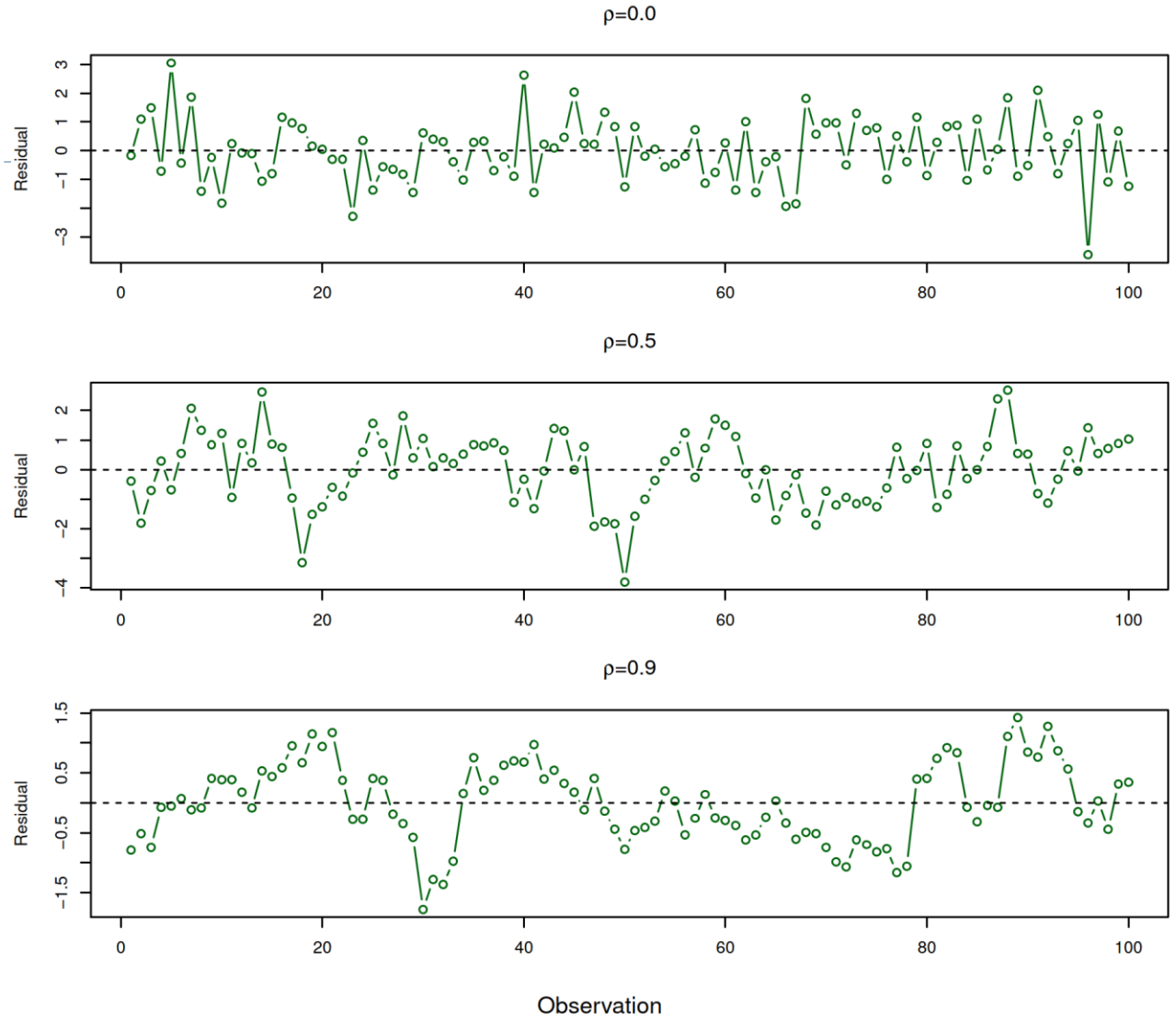
$\hat{y}_i$

2. Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, are uncorrelated. What does this mean?
  - For instance, if the errors are uncorrelated, then the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}
  - If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be
  - As an extreme example, suppose we accidentally doubled our data, leading to observations and error terms identical in pairs. If we ignored this, our standard error calculations would be as if we had a sample of size $2n$, when in fact we have only n samples. Our estimated parameters would be the same for the $2n$ samples as for the n samples, but the confidence intervals would be narrower by a factor of $\sqrt{2}$!

2. Correlation of Error Terms

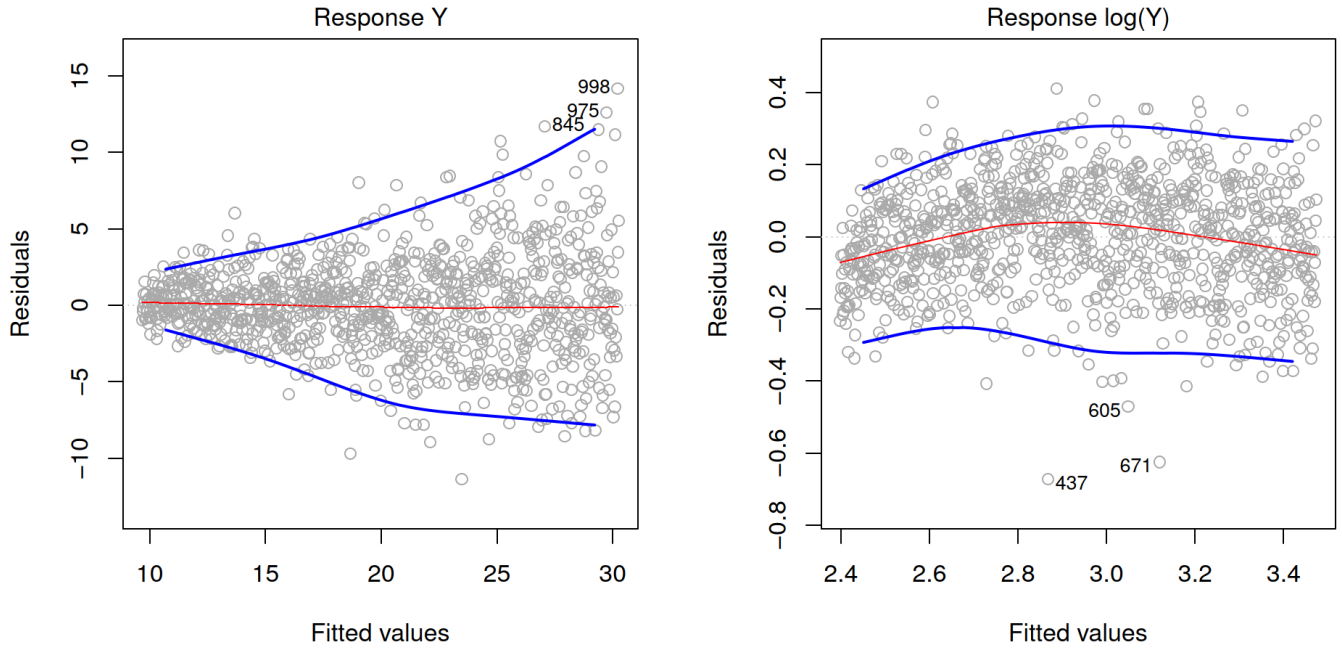
- In the top panel, we see the residuals from a linear regression fit to data generated with *uncorrelated errors*
- The residuals in the bottom panel shows a clear pattern in the residuals—adjacent residuals tend to take on similar values
- Finally, the center panel illustrates a more moderate case in which the residuals had a correlation of 0.5. There is still evidence of tracking, but the pattern is less clear



Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time point

3. Non-constant Variance of Error Terms

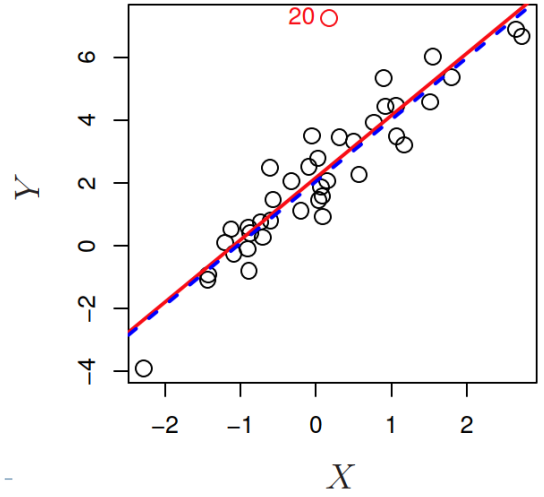
- It is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response. One can identify non-constant variances in the errors, or *heteroscedasticity*, from the presence of a funnel shape in the residual plot

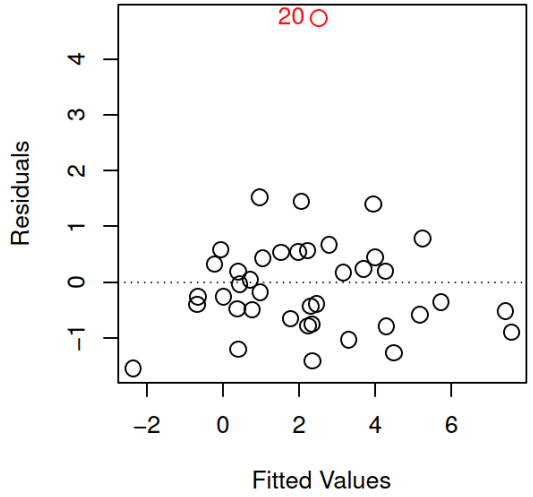


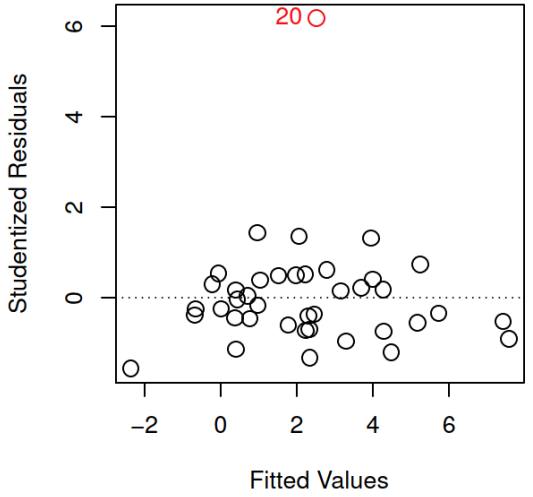
The right-hand panel displays the residual plot after transforming the response using $\log Y$

4. Outliers

- An outlier is a point for which y_i is far from the value predicted by the model
  - Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection
  - The red solid line is the least squares regression fit, while the blue dashed line is the least squares fit after removal of the outlier
  - Observations whose studentized residuals are greater than 3 in absolute value are possible outliers







Studentized Residuals

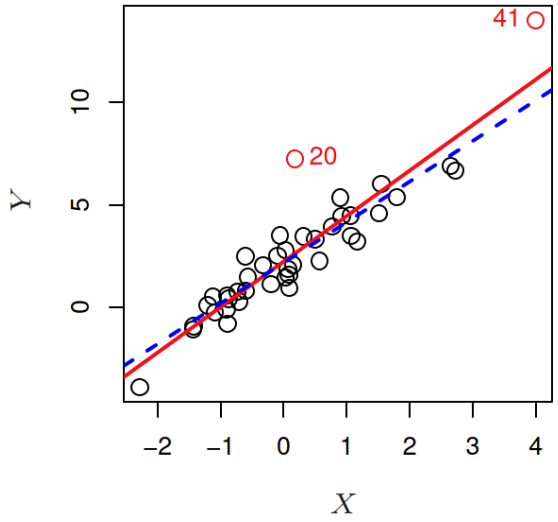
$$t_i = \frac{e_i}{SE(e_i)} = e_i/(\hat{\sigma}\sqrt{1-h_i})$$

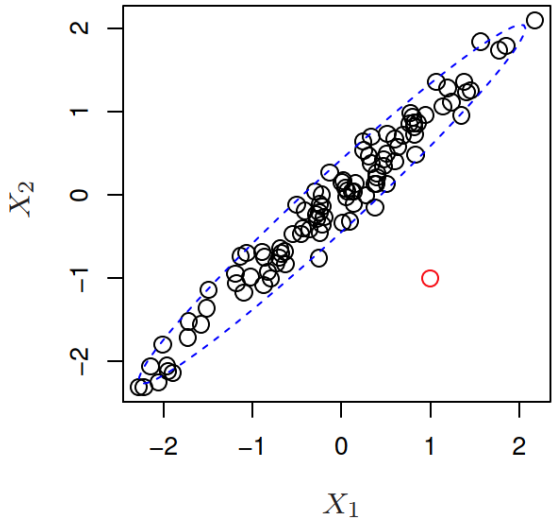
4. Outliers

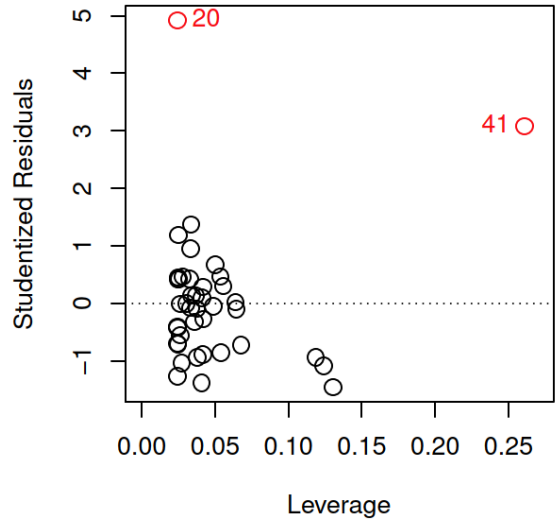
- It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit. However, even if an outlier does not have much effect on the least squares fit, it can cause other problems
  - For instance, in this example, the RSE is 1.09 when the outlier is included in the regression, but it is only 0.77 when the outlier is removed
  - Since the RSE (σ) is used to compute all confidence intervals and p-values, such a dramatic increase caused by a single data point can have implications for the interpretation of the fit
- If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation
  - However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor

5. High Leverage Points

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i
  - In contrast, observations with high leverage have an unusual value for x_i
  - In order to quantify an observation's leverage, we compute the *leverage statistic*. A large value of this statistic indicates an observation with high leverage







5. High Leverage Points

- For a simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

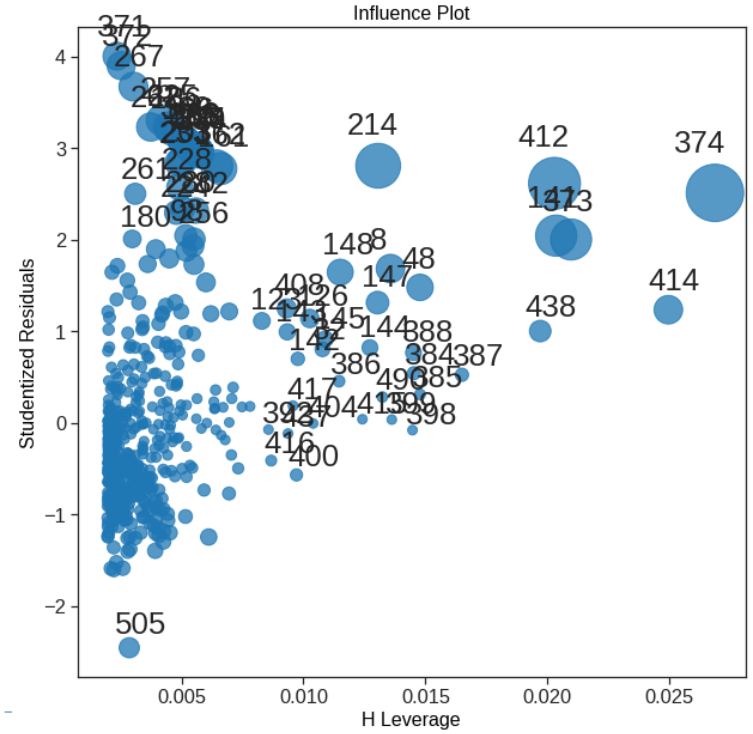
  - The leverage statistics is always between $1/n$ and 1, and the average leverage for all the observations is always equal to $(p+1)/n$
  - So if a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage
- A value whose absence would significantly change the regression equation is termed an *influential observation*
  - Although an influential point will typically have high leverage, a high leverage point is not necessarily an influential point
  - Comparison between influential and high leverage point

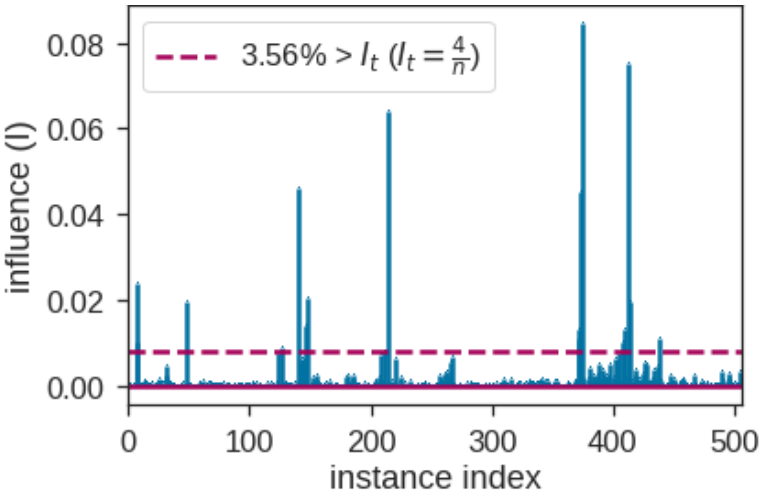
5. High Leverage Points

- *Cook's distance* is a commonly used estimate of the influence of a data point

$$\frac{1}{p+1} t_i^2 \frac{h_i}{1-h_i}$$

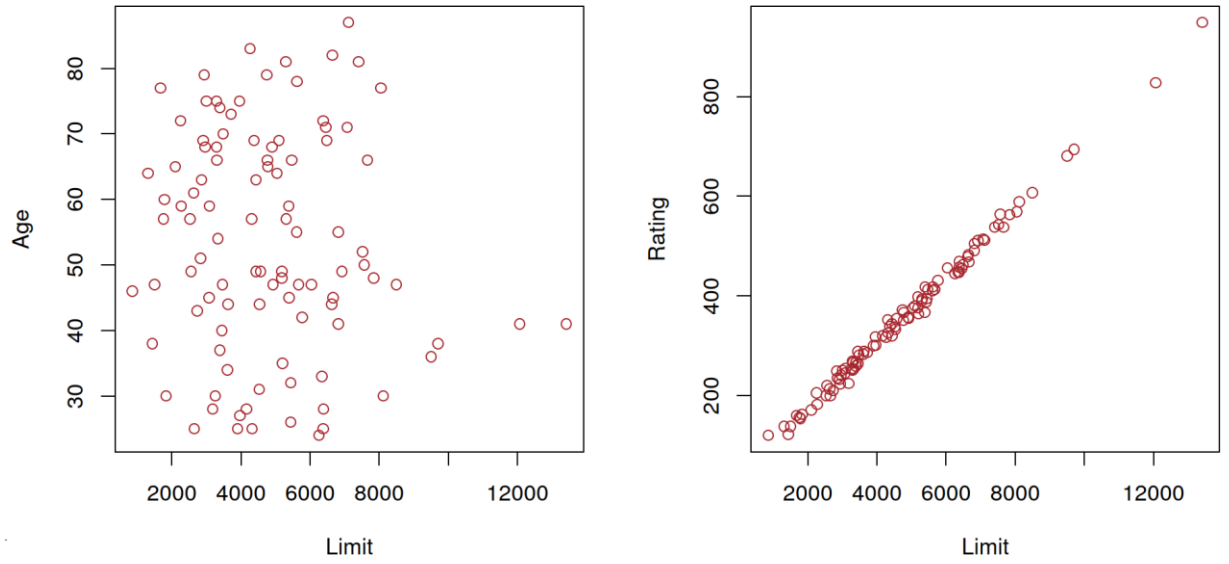
- An influential plot can be used to analyze data points
- How to deal with influential points?
  - It indicates data points that are particularly worth checking for validity
  - It indicates regions of the design space where it would be good to be able to obtain more data points





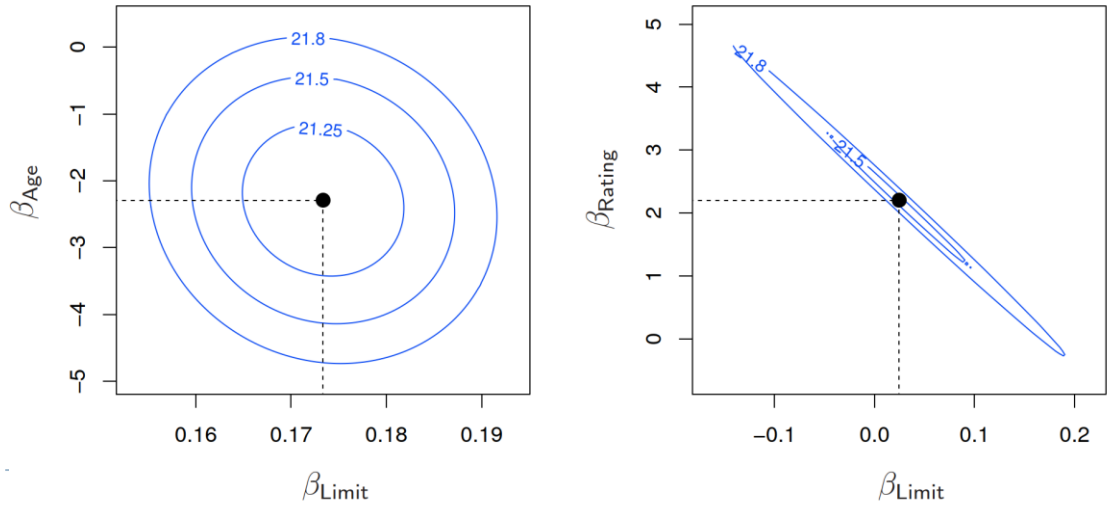
6. Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another
  - The concept of collinearity is illustrated below using the Credit data set
  - In the left-hand panel, the limit and age appear to have no obvious relationship
  - In contrast, in the right-hand panel, the predictors limit and rating are very highly correlated with each other, and we say that they are collinear



6. Collinearity

- Figure below illustrates some of the difficulties that can result from collinearity
  - The left-hand panel is a contour plot of the RSS associated with different possible coefficient estimates for the regression of balance on limit and age
  - Each ellipse represents a set of coefficients that correspond to the same RSS, with ellipses nearest to the center taking on the lowest values of RSS
  - The black dots and associated dashed lines represent the coefficient estimates that result in the smallest possible RSS—in other words, these are the least squares estimates



6. Collinearity

- The right-hand panel displays contour plots of the RSS associated with possible coefficient estimates for the regression of balance onto limit and rating
  - The contours run along a narrow valley; there is a broad range of values for the coefficient estimates that result in equal values for RSS
  - Hence a small change in the data could cause the pair of coefficient values that yield the smallest RSS—that is, the least squares estimates—to move anywhere along this valley
  - This results in a great deal of uncertainty in the coefficient estimates and thus reduce the *power* of the hypothesis test
- Table 3.11 compares the coefficient estimates obtained from two separate multiple regression models
  - In the first regression, both age and limit are highly significant with very small p-values
  - In the second, the collinearity between limit and rating has caused the standard error for the limit coefficient estimate to increase by a factor of 12 and the p-value to increase to 0.701

6. Collinearity

- In other words, the importance of the limit variable has been masked due to the presence of collinearity
  - A simple way is to use a correlation matrix to detect collinearity. Unfortunately, it is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

6. Collinearity

- A better way to assess *multicollinearity* is to compute the *variance inflation factor (VIF)* using the formula

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

  where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors

- In the Credit data, a regression of balance on age, rating, and limit indicates that the predictors have VIF values of 1.01, 160.67, and 160.59
  - As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problem
- When faced with the problem of collinearity
  1. The first is to drop one of the problematic variables from the regression since it is redundant
  2. The second solution is to combine the collinear variables together into a single predictor

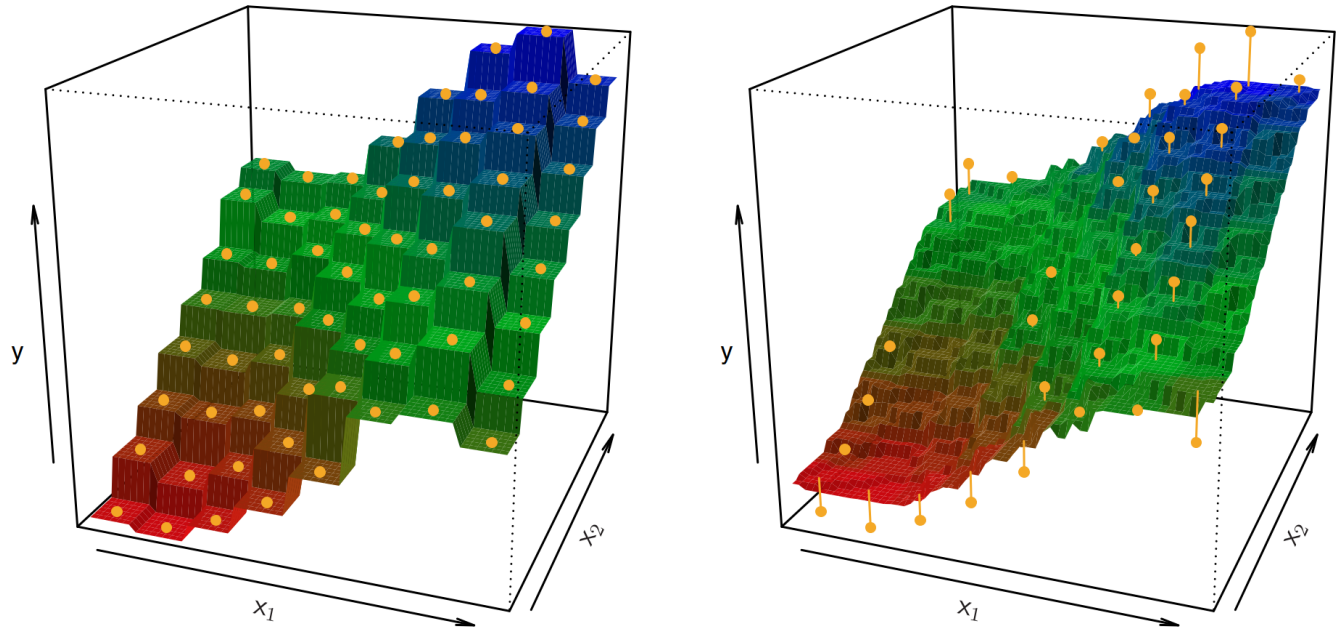
Comparison of Linear Regression with K-Nearest Neighbors

- The non-parametric methods do not explicitly assume a parametric form for $f(X)$, and thereby provide an alternative and more flexible approach for performing regression
  - Here we consider one of the simplest and best-known non-parametric methods, *K-nearest neighbors regression* (KNN regression)
  - Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by N_0
  - It then estimates $f(x_0)$ using the average of all the training responses in N_0. In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

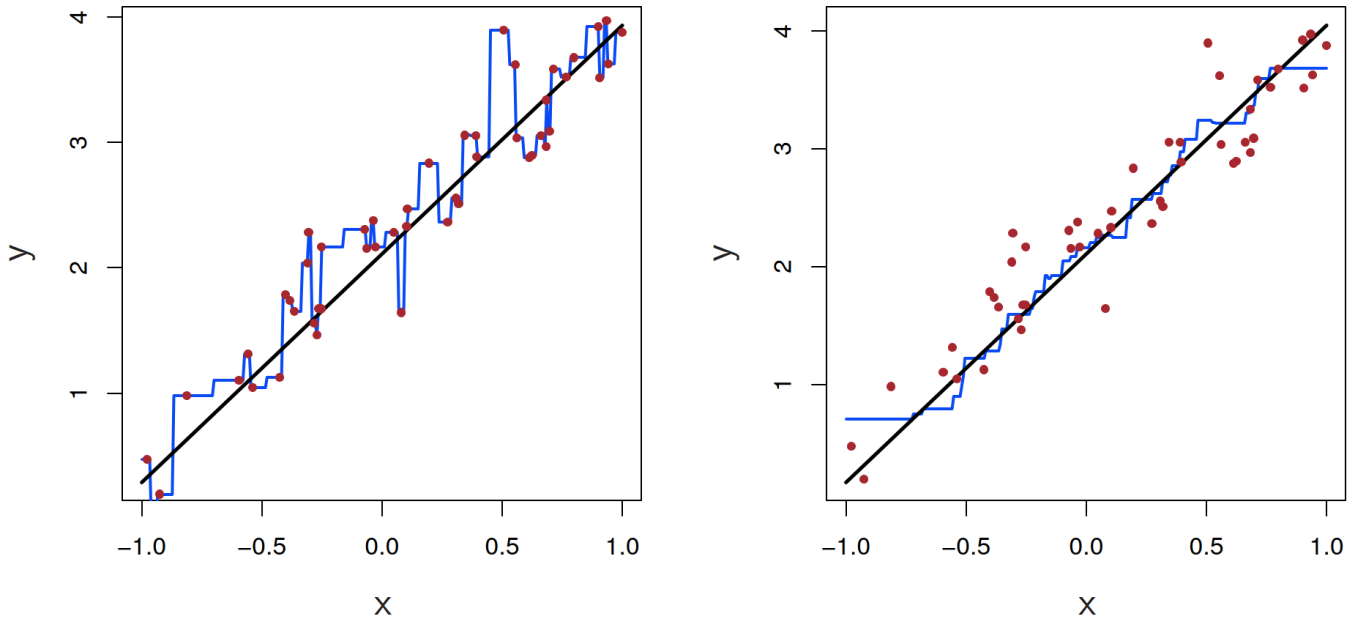
Comparison of Linear Regression with K-Nearest Neighbors

- Figure below illustrates two KNN fits on a data set with $p = 2$ predictors
  - The fit with $K = 1$ is shown in the left-hand panel, while the right-hand panel corresponds to $K = 9$
- In general, the optimal value for K will depend on the bias-variance tradeoff
  - A small value for K provides the most flexible fit, which will have low bias but high variance



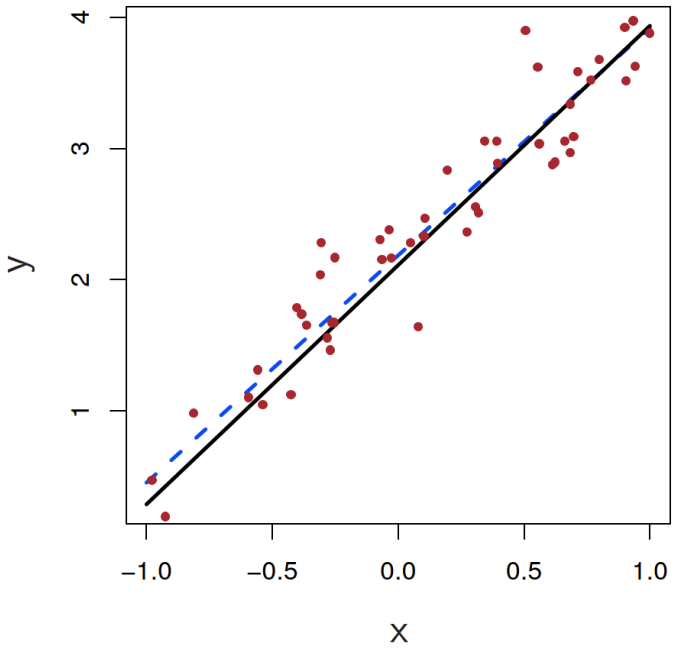
Comparison of Linear Regression with K-Nearest Neighbors

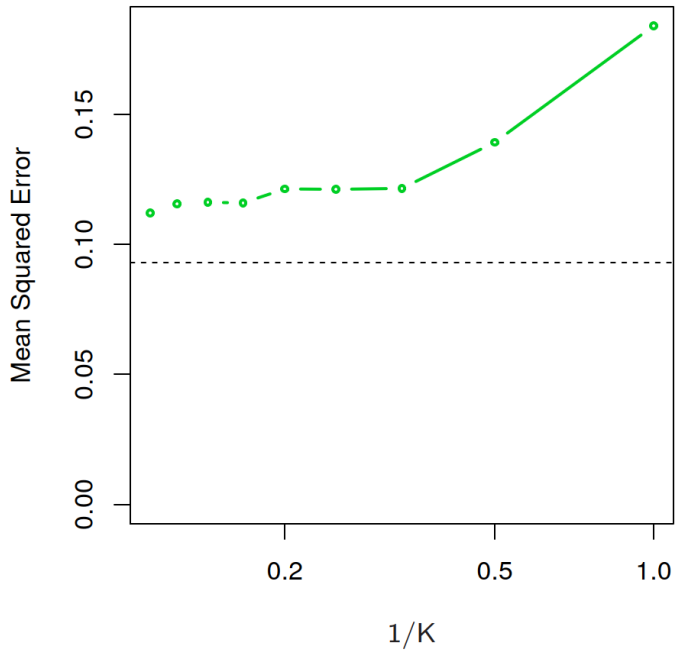
- The black solid lines represent $f(X)$, while the blue curves correspond to the KNN fits using $K = 1$ and $K = 9$
  - In this case, the $K = 1$ predictions are far too variable, while the smoother $K = 9$ fit is much closer to $f(X)$
  - However, since the true relationship is linear, it is hard for a non-parametric approach to compete with linear regression: a non-parametric approach incurs a cost in variance that is not offset by a reduction in bias



Comparison of Linear Regression with K-Nearest Neighbors

- The blue dashed line in the left-hand panel of Figure represents the linear regression fit to the same data
  - In the right hand panel, the green solid line, plotted as a function of $1/K$, represents the test set mean squared error (MSE) for KNN
  - The KNN errors are well above the black dashed line, which is the test MSE for linear regression. When the value of K is large, then KNN performs only a little worse than least squares regression in terms of MSE

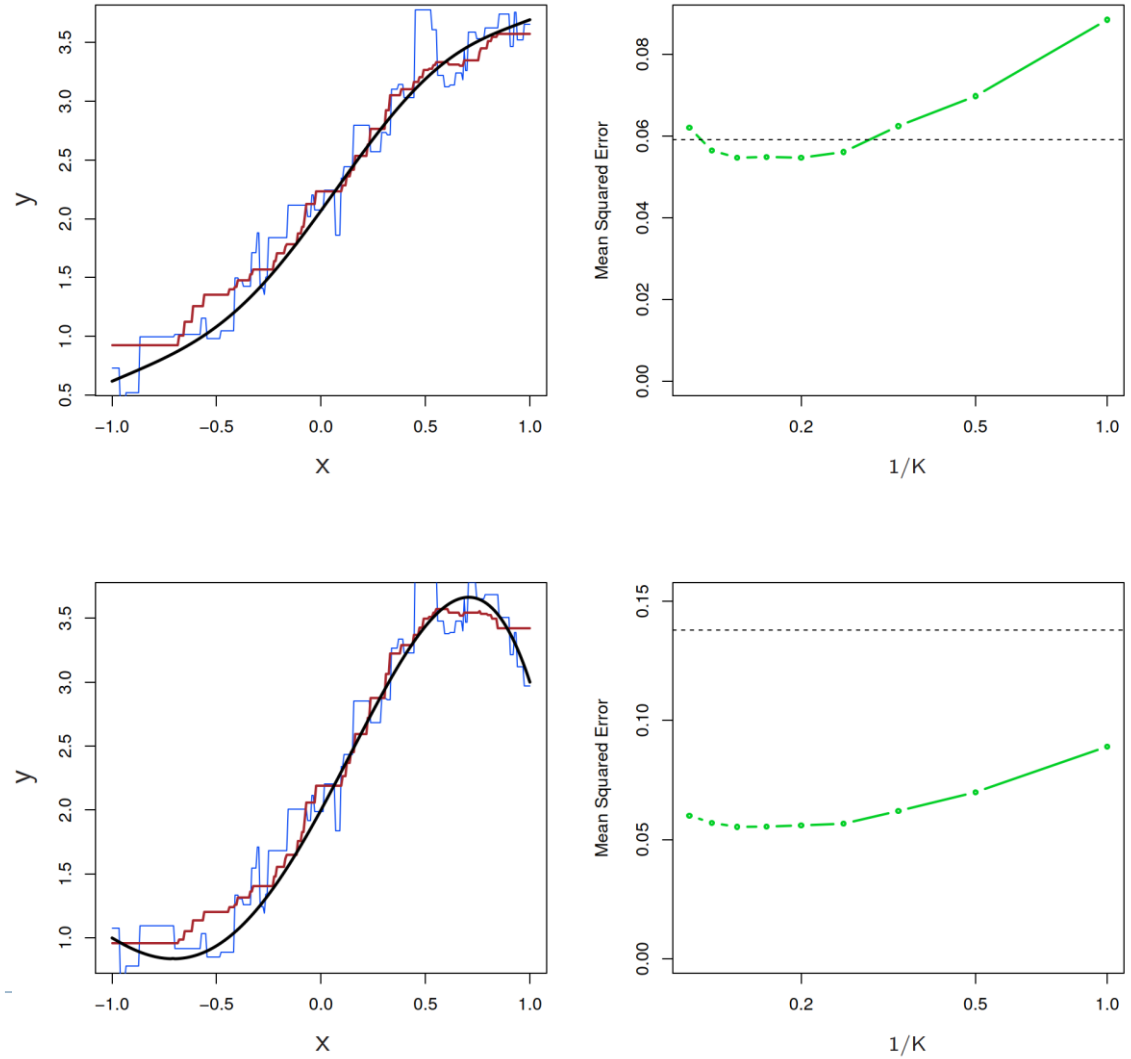




$$MSE = \frac{RSS}{n}$$

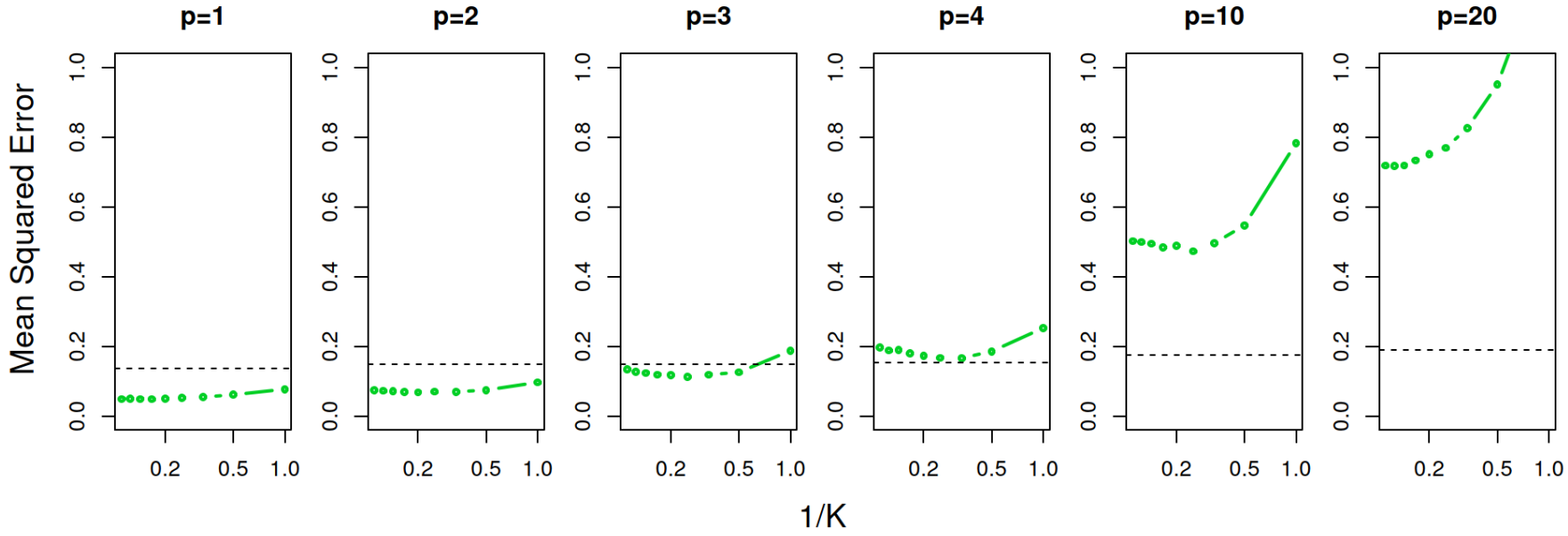
Comparison of Linear Regression with K-Nearest Neighbors

- In practice, the true relationship between X and Y is rarely exactly linear
  - In the top row, the true relationship is close to linear
  - The second row illustrates a more substantial deviation from linearity. In this situation, KNN substantially outperforms linear regression for all values of *K*
  - Note that as the extent of non-linearity increases, there is little change in the test set MSE for the non-parametric KNN method, but there is a large increase in the test set MSE of linear regression



Comparison of Linear Regression with K-Nearest Neighbors

- Figure below considers the same strongly non-linear situation except that we have added additional noise predictors that are not associated with the response
  - In fact, the increase in dimension has only caused a small deterioration in the linear regression test set MSE, but it has caused more than a ten-fold increase in the MSE for KNN
  - This decrease in performance as the dimension increases is a common problem for KNN, and results from the fact that in higher dimensions there is effectively a reduction in sample size



Generalizations of the Linear Model

- In much of the rest of this course, we discuss methods that expand the scope of linear models and how they are fit:
  - Classification problems: logistic regression, support vector machines
  - Non-linearity: kernel smoothing, splines and generalized additive models
  - Interactions: Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
  - Regularized fitting: Ridge regression and lasso

Appendix

Reference

- About the proof of linear regression
  - https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares
  - https://en.wikipedia.org/wiki/Simple_linear_regression
- About the concept of linear regression
  - https://online.stat.psu.edu/stat501/lesson/3/3.3
  - https://online.stat.psu.edu/stat415/lesson/8/8.1
  - https://stats.stackexchange.com/questions/85560/shape-of-confidence-interval-for-predicted-values-in-linear-regression

The woes of (interpreting) regression coefficients

- "Data Analysis and Regression" Mosteller and Tukey 1977
  - A regression coefficient β_j estimates the expected change in Y per unit change in X_j, with all other predictors held fixed. But predictors usually change together!
  - Example: Y total amount of change in your pocket; X_1 =# of coins; X_2 = # of pennies, nickels and dimes. By itself, regression coefficient of Y on X_2 will be > 0. But how about with X_1 in model?
  - Y = number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $Y = b_0 + 0.50W - 0.10H$. How do we interpret $\hat{\beta}_2 < 0$?

Two quotes by famous Statisticians

- “Essentially, all models are wrong, but some are useful” - George Box

- “The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively” - Fred Mosteller and John Tukey, paraphrasing George Box

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated - a balanced design:
  - Each coefficient can be estimated and tested separately
  - Interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible

- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous - when X_j changes, everything else changes

- *Claims of causality* should be avoided for observational data