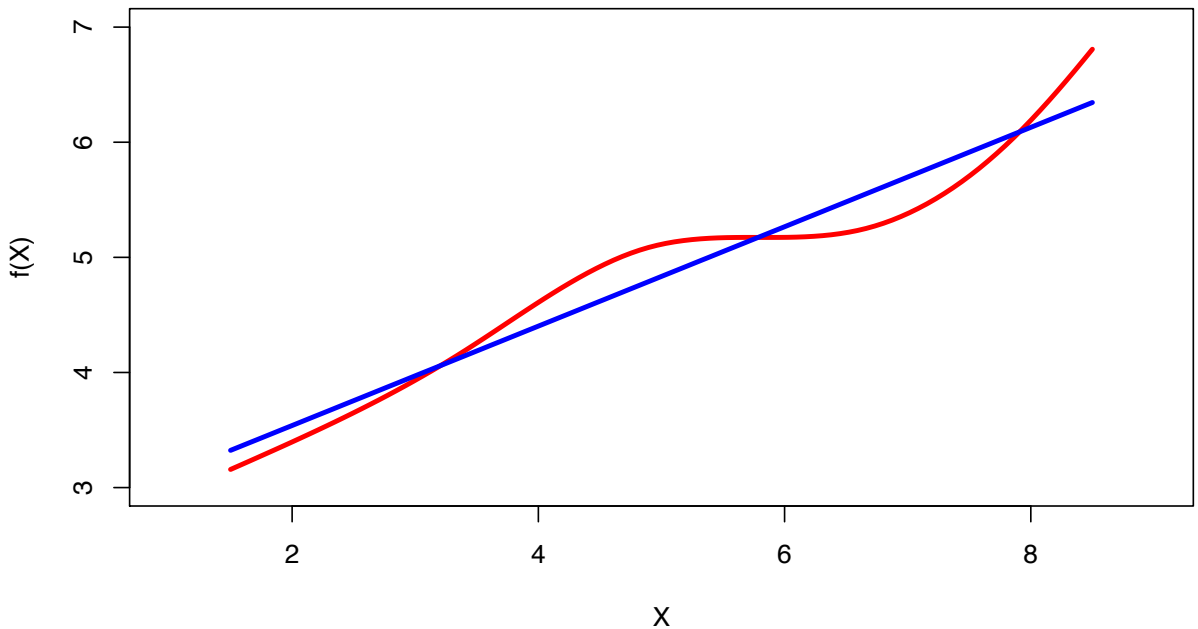
3.1 Simple Linear Regression

https://youtu.be/PsE9UqoWtS4

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 shown on the next slide.

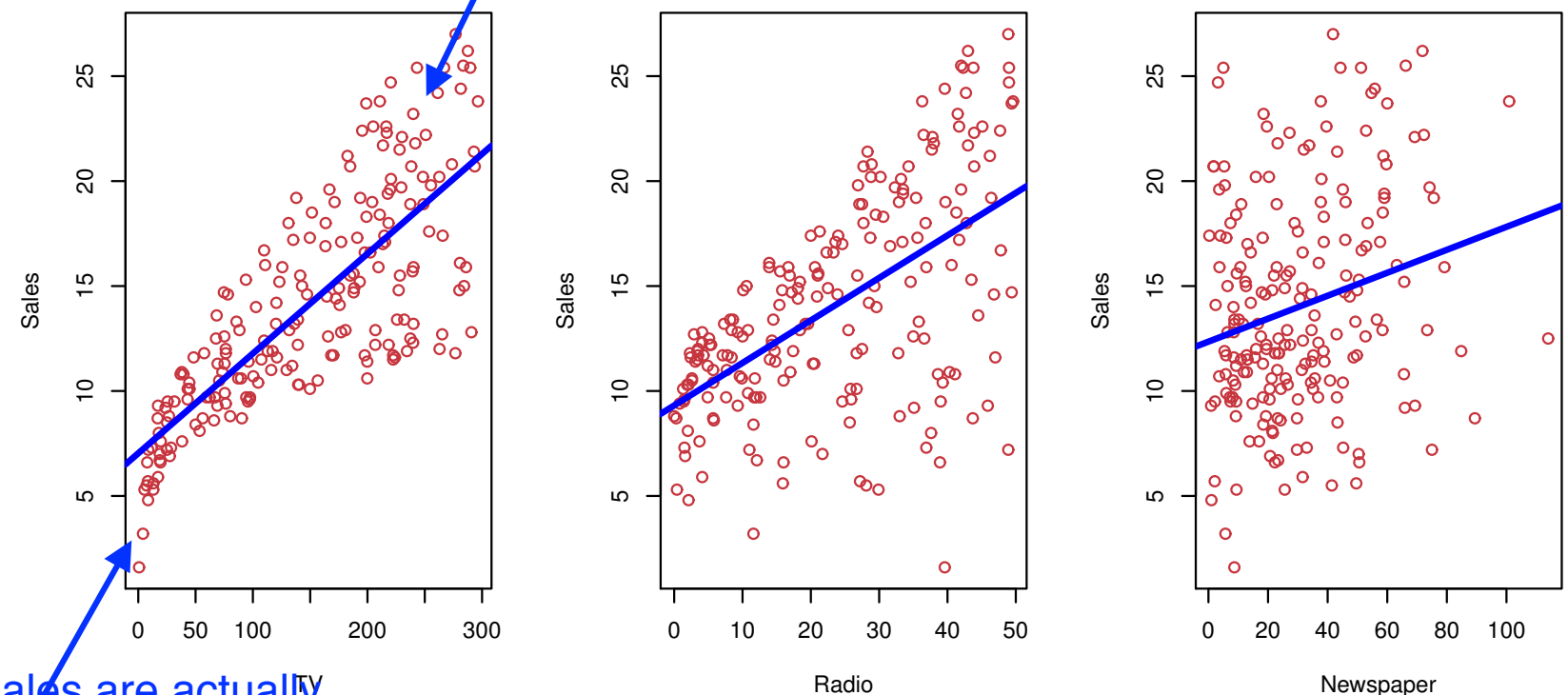
Questions we might ask:

- 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?

  In other words, do the media work on their own in a certain way, or do they work in combination?

Advertising data

the amount of noise around the curve/around the line, is quite large.



the sales are actually lower than expected,

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.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ 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

- 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

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

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

least squares estimates : (These are the ones that minimize the sum of squares.)

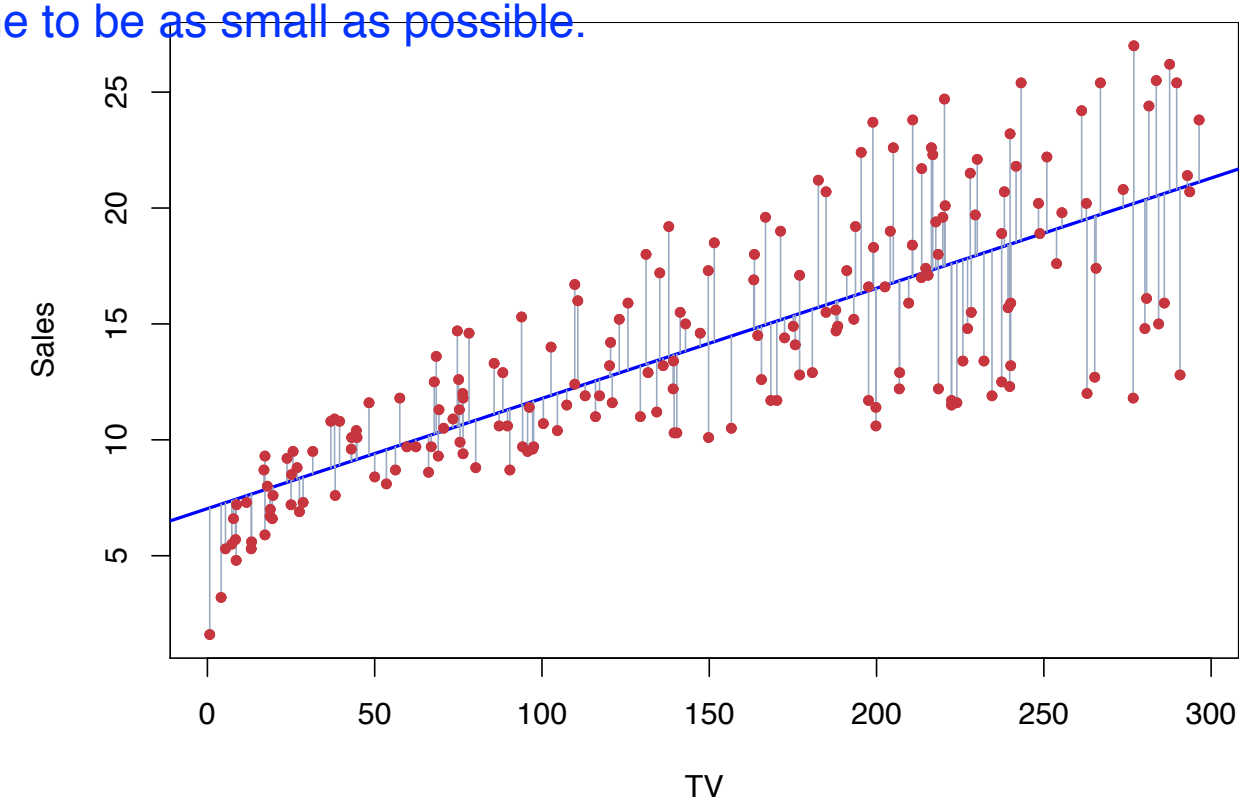
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

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

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ are the sample means.

Example: advertising data

But I want the total squared distance of all points to the line to be as small as possible.



least squares line

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 somewhat deficient in the left of the plot.

how precise are those estimates ?

Assessing the Accuracy of the Coefficient Estimates

(1) the standard error of the slope is bigger if my noise variance is bigger.
(2) the more spread out the x's, the more precise the slope is.
so maybe in an experiment where you can design, you should pick your predictor values, the x's, as spread out as possible in order to get the slopes estimated as precisely as possible.

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

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

where $\sigma^2 = \mathrm{Var}(\epsilon)$ sigma squared is the noise, the variance of the errors around the line.

- 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

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

if errors are normally distributed, which we typically assume, approximately, this will contain the true value, the true slope, with probability 0.95.

check Hypothesis testing.pptx

Confidence intervals — continued

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

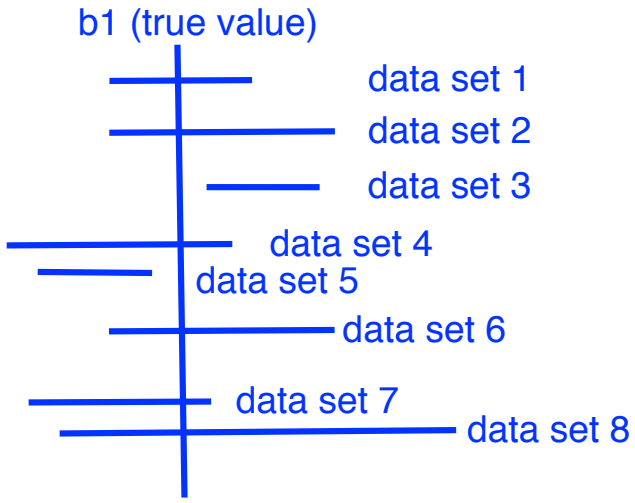
$$\Big[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\Big]$$

lower point upper point

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 $[0.042, 0.053]$

In other words, having TV advertising does have a positive effect on sales, as one would expect.



what the theory tells us is that if I form, say, 100 confidence intervals, 100 of these brackets, 95% of the time, they will contain the true value. The other 5% of the time, they will not contain the true value.

it's a test of a certain value of a parameter.

check Hypothesis testing.pptx

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

$$H_0 : \quad \text{There is no relationship between } X \text{ and } Y$$

versus the *alternative hypothesis*

$$H_A : \quad \text{There is some relationship between } X \text{ and } Y.$$

- Mathematically, this corresponds to testing

here it's a test "is the parameter (slope, beta 1) 0 ?"

$$H_0 : \beta_1 = 0$$

versus

The alternative hypothesis is that there is some relationship between X and Y. In other words, Beta 1 is not 0.

$$H_A : \beta_1 \neq 0,$$

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

3.2 Hypothesis Testing and Confidence Intervals

https://youtu.be/J6AdoiNUyWI

Hypothesis testing — continued

It's basically you look this up in a table or, nowadays software will compute it for you.

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

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

- 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*.

In any case, you ask the computer to compute the p-value based on this statistic. p-value is the probability of getting the value of t at least as large as you got in absolute value.

Results for the advertising data

check Hypothesis testing.pptx

	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

So how do we interpret this? It says the chance of seeing this data, under the assumption that the null hypothesis (there's no effect of TV advertising on sales) is less than 10 to the minus 4. So it's very unlikely to have seen this data. It's possible, but very unlikely under the assumption that TV advertising has no effect. Our conclusion, therefore, is that TV advertising has an effect on sales--

http://blog.minitab.com/blog/understanding-statistics/three-things-the-p-value-cant-tell-you-about-your-hypothesis-test
In hypothesis testing, when your p-value is less than the alpha level you selected (typically 0.05), you'd reject the null hypothesis in favor of the alternative hypothesis.

for hypothesis test and confidence intervals, there's actually a one-to-one correspondence, they're doing equivalent things. To be more precise, if hypothesis test fails (in other words, if we reject the null hypothesis and conclude that Beta 1 is not 0, as we did for TV advertising), correspondingly the confidence interval constructed for that data for the parameter will not contain 0. Conversely, if the hypothesis test does not reject, so we cannot conclude that TV advertising has an effect. Its slope may be 0. The confidence interval for that parameter will contain 0. So really, the confidence interval is also doing hypothesis testing for you.But it's also telling you how big the effect is. So it's always good to compute confidence intervals as well as do hypothesis test.

So for example, here we see the beta 1 interval [0.042, 0.053] doesn't contain 0. . Furthermore, we see that a lower limit on the effect of TV advertising is 0.042, which we can interpret as -- for each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units. So this tells us not only is the effect 0 or not, but how big is the effect likely to be

How about the overall fit of the model, the accuracy of the model?

Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

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

  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

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

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

if we didn't fit a model at all and just use the mean of sales as the prediction (no predictor model), that's the simplest prediction you can imagine. TSS would be our error. And now, the residual sum of squares of the fitted model is RSS, and since we've done least squares, we've optimized over the parameters. We know that RSS will be less than TSS.

no predictor/mean/null

- 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:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

The R squared is 0.61, in other words, using TV budget, We reduced the variance in sales by 61%.

In business/finance and some kind of physical sciences, we see R squares like this, In medicine, you might see an R square of 5% and you might get excited.

3.3 Multiple Linear Regression

https://youtu.be/1hbCJyM9ccs

Multiple Linear Regression

we have more than 1 predictors

- 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

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Interpreting regression coefficients

no correlation between variables

- 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.

if there is correlation between variables

then it's not reasonable to assume "that all the other variables stay fixed"

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

We can't really say that one predictor causes the outcome when there's predictors in the system that are correlated with that given predictor, so it becomes challenges to discuss causality, and we are gonna avoid that.

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?

  X1, X2 are highly correlated, we cannot really "with all other predictors held fixed" ...
- $Y =$ number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $\hat{Y} = b_0 + .50W - .10H$. How do we interpret $\hat{\beta}_2 < 0$?

  it seems to say it's better to be short to make more tackles, but it doesn't make sense. How to interpret it then ?

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

In other words, if you want to make a causal statement about a predictor for an outcome, you actually have to be able to take the system and perturb that particular predictor, keeping the other ones fixed. That will allow you to make a causal statement about a variable like xj and its effect on the outcome. you cannot merely observe it.

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

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

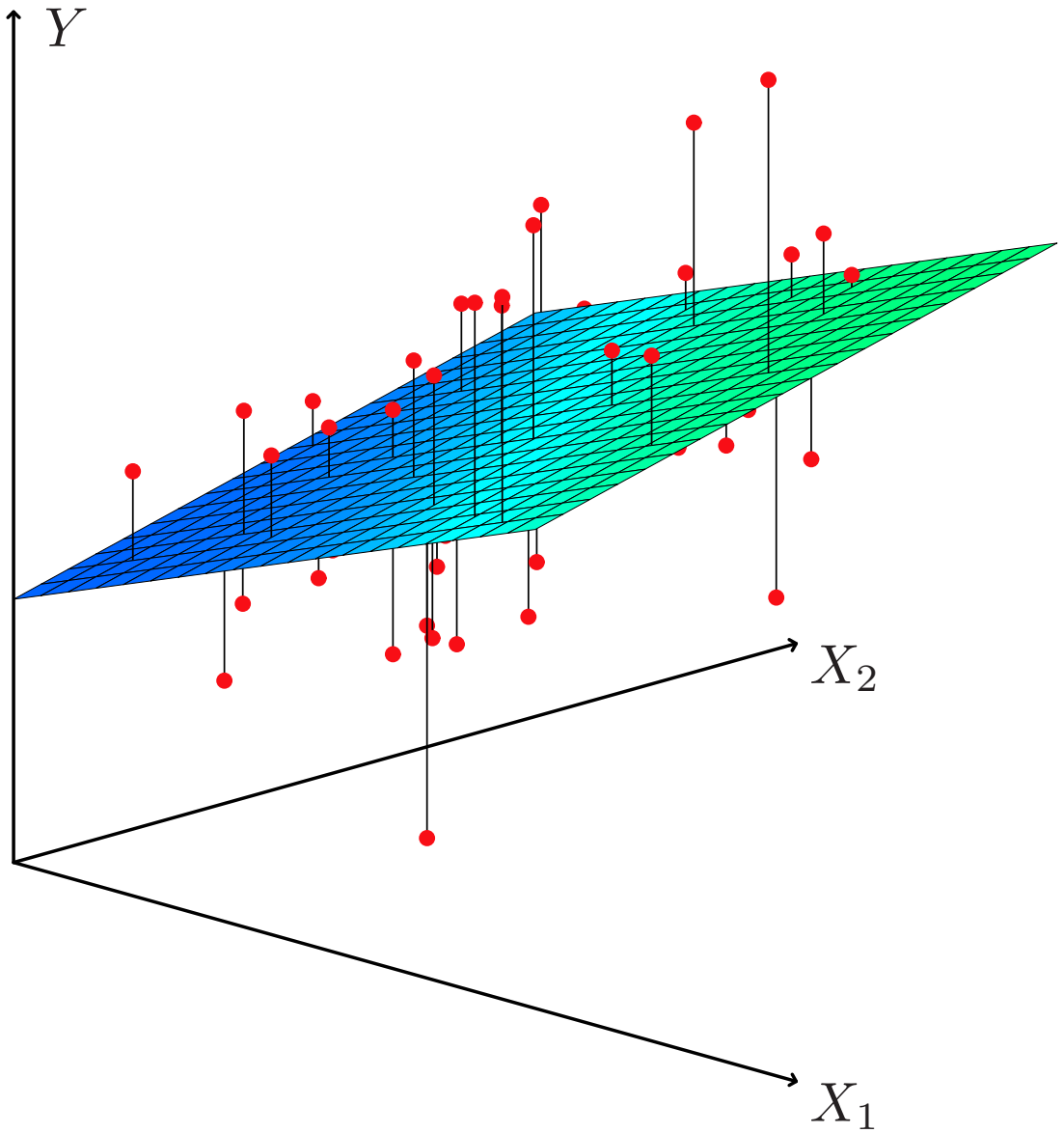
- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

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

there is a formula for these coefficients

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.

multiple regression to fit a hyperplane to these points to minimize the distance between points and the hyperplane



Results for advertising data

for TV and radio, it the presence of other 2 predictors fixed, the effect is significant, but newspaper is not.

	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

for newspaper, p-value is large, it's not evidence against the null hypothesis, (null: coeff. = 0)

Correlations:

	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

radio and newspaper has correlation of 0.3541, so whats likely happens here is that any effect of newspaper has been soaped up by radio because they are correlated at 0.35. so with radio in the model, newspaper is no longer needed.It doesn't improve the prediction given we've measured the radio advertising. On the other hand, TV and radio are uncorrelated, so their effects are somewhat complimentary.

3.4 Some important questions

https://youtu.be/3T6RXmIHbJ4

Some important questions

these are all things we can answer from a model

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

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

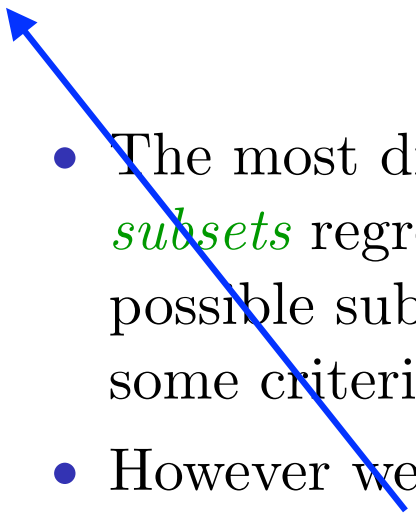
Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

R^2 ~ 0.6 if only consider TV (computed previously), now by adding 2 more predictors we increases it to 0.897

TSS: use “no predictor” model (if we just use mean to predict)
TSS - RSS = drop of training error, proportional to “variance explained”
variance explained : R^2 ~ 0.897 => we reduces the variance of sales around its mean by ~90% by using these 3 predictors
p=number of number of parameters = 3 here
n-p-1 : degrees of freedom

ex:
if p = 2 (=> x1, x2)
y = b0
y = b0 + b1*x1
y = b0 + b2*x2
y = b0 + b1*x1 + b2*x2

Deciding on the important variables



- 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!
  Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

Forward selection

And the intercept is the mean of y with no other variables

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

you search through all the single-variable models and pick the best one.And now you search through the remaining (p-1) variables again and find out which variable should be added to the variable you've already picked to best improve the residual sum of squares. And you continue until some stopping rule is satisfied--

So in a similar fashion, and if p is not too large, you can start from the other end.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
- 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.

Model selection — continued

- Later we discuss more systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.

- These include *Mallow's C_p*, *Akaike information criterion (AIC)*, *Bayesian information criterion (BIC)*, *adjusted R^2* and *Cross-validation (CV)*.

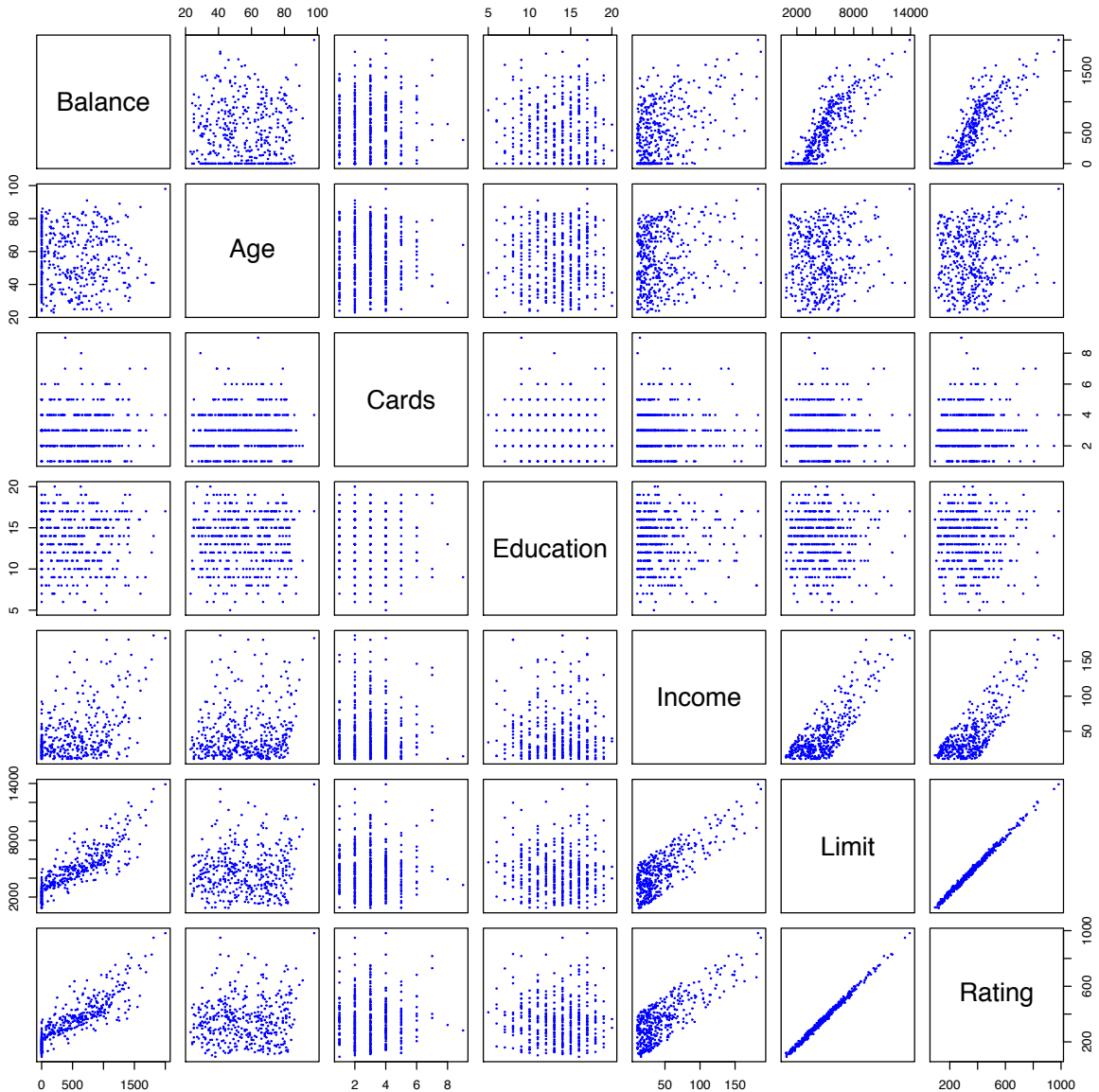
Other Considerations in the Regression Model

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: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

Credit Card Data



Qualitative Predictors — continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Intrepretation?

Credit card data — continued

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

large p value => the coefficient is not signifiant

So contrary to popular wisdom, females don't generally have a higher credit card balance than males. The number 19.73 is slightly higher, but it's not significant.

Qualitative predictors with more than two levels

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

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

and the second could be

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

Qualitative predictors with more than two levels — continued.

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

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

And if it's got k levels, you'll make k minus 1 dummy variables to represent each of those categories.

Results for ethnicity

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

the baseline is AA, its comparing Asian with AA, and its not significant (large p value)

the baseline is AA, its comparing Caucasian with AA, and its not significant (large p value)

Now, it turns out the choice of the baseline does not affect the fit of the model.
The residual sum of squares would be the same no matter which category you chose as the baseline. But the contrasts would change, because picking the baseline determines which contrasts you make. And so the p values potentially would change as you change the baseline.

3.5 Extensions of the linear model

https://youtu.be/IFzVxLv0TKQ

So that's one extension of the linear model. Another extensions are interactions and non-linearity.

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

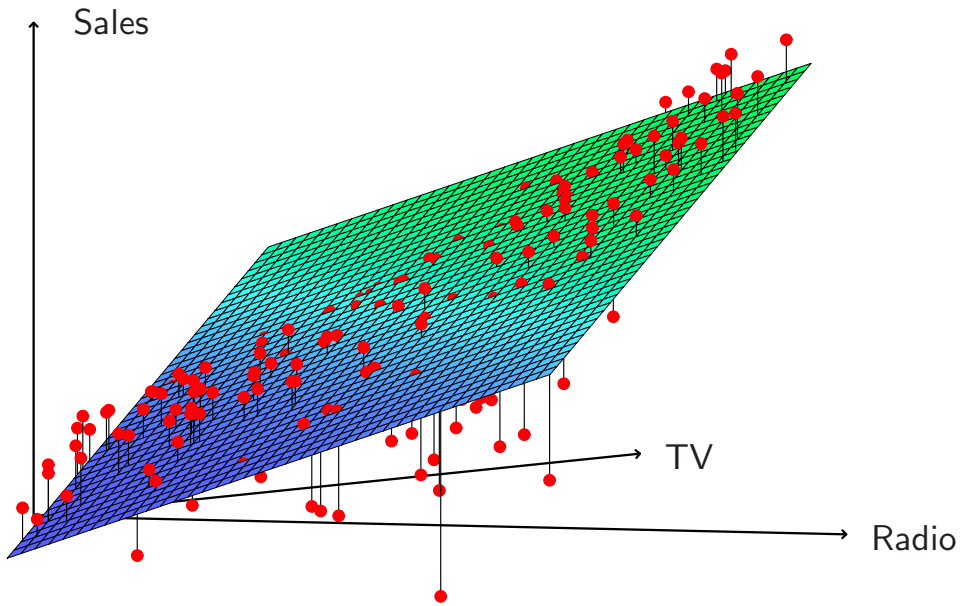
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

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`.

Interactions — continued

- 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.
- In this situation, 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.

Interaction in the Advertising data?



When levels of either `TV` or `radio` are low, then the true `sales` are lower than predicted by the linear model.
But when advertising is split between the two media, then the model tends to underestimate `sales`.

Modelling interactions — Advertising data

Model takes the form

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + \boxed{(\beta_1 + \beta_3 \times \texttt{radio})} \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

coefficient of TV, which had been originally beta_1, is now modified as a function of radio. So as the values of radio changes, the coefficient of TV changes by amount beta_3 * radio.

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	b0 6.7502	0.248	27.23	< 0.0001
TV	b1 0.0191	0.002	12.70	< 0.0001
radio	b2 0.0289	0.009	3.24	0.0014
TV×radio	b3 0.0011	0.000	20.73	< 0.0001

the interaction is significant !! which is consistent with the previous picture

Interpretation

- The results in this table suggests that interactions are important.
- The p-value for the interaction term `TV`×`radio` 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.

Interpretation — continued

- 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 $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

Hierarchy

- 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.

Hierarchy — continued

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

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

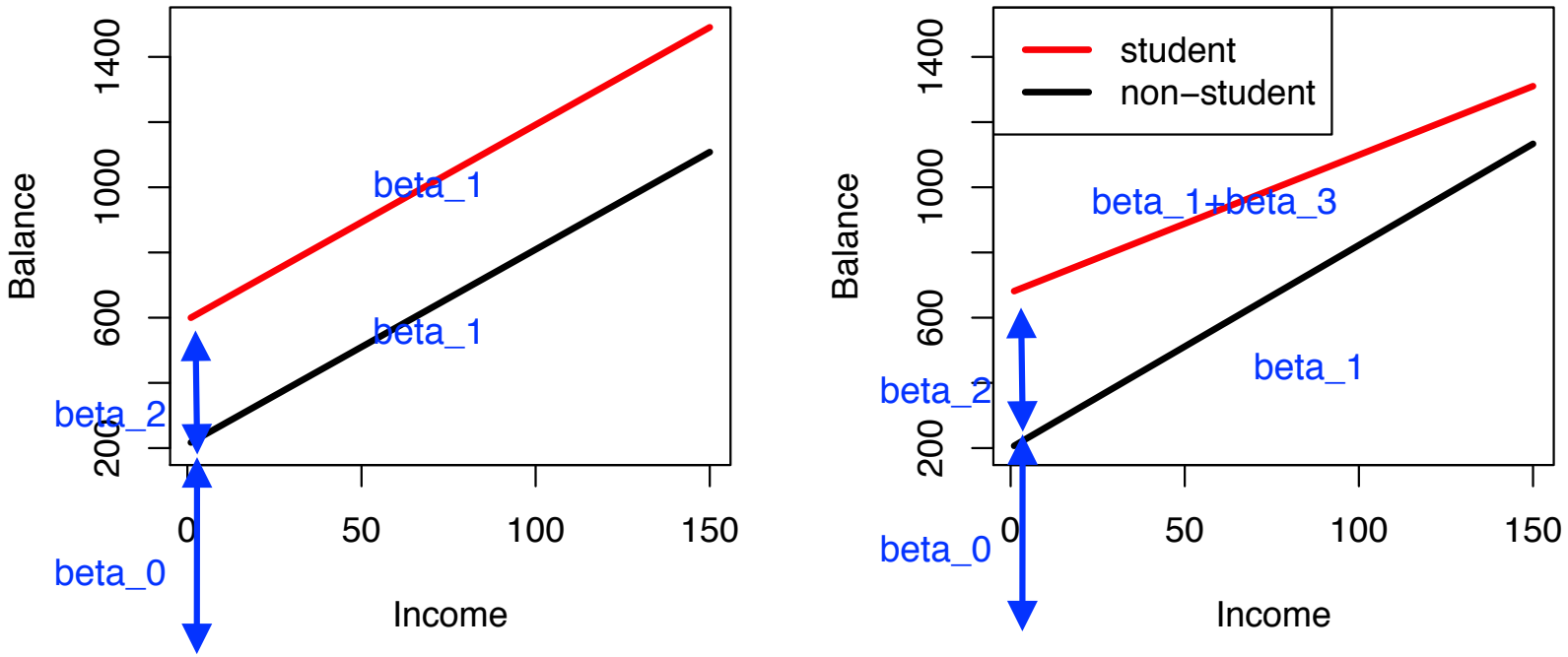
we have dummy variables for students

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}$$

we can think of this as having a common slope in income, but a different intercept depending on whether the person is a student or not.

With interactions, it takes the form

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}$$

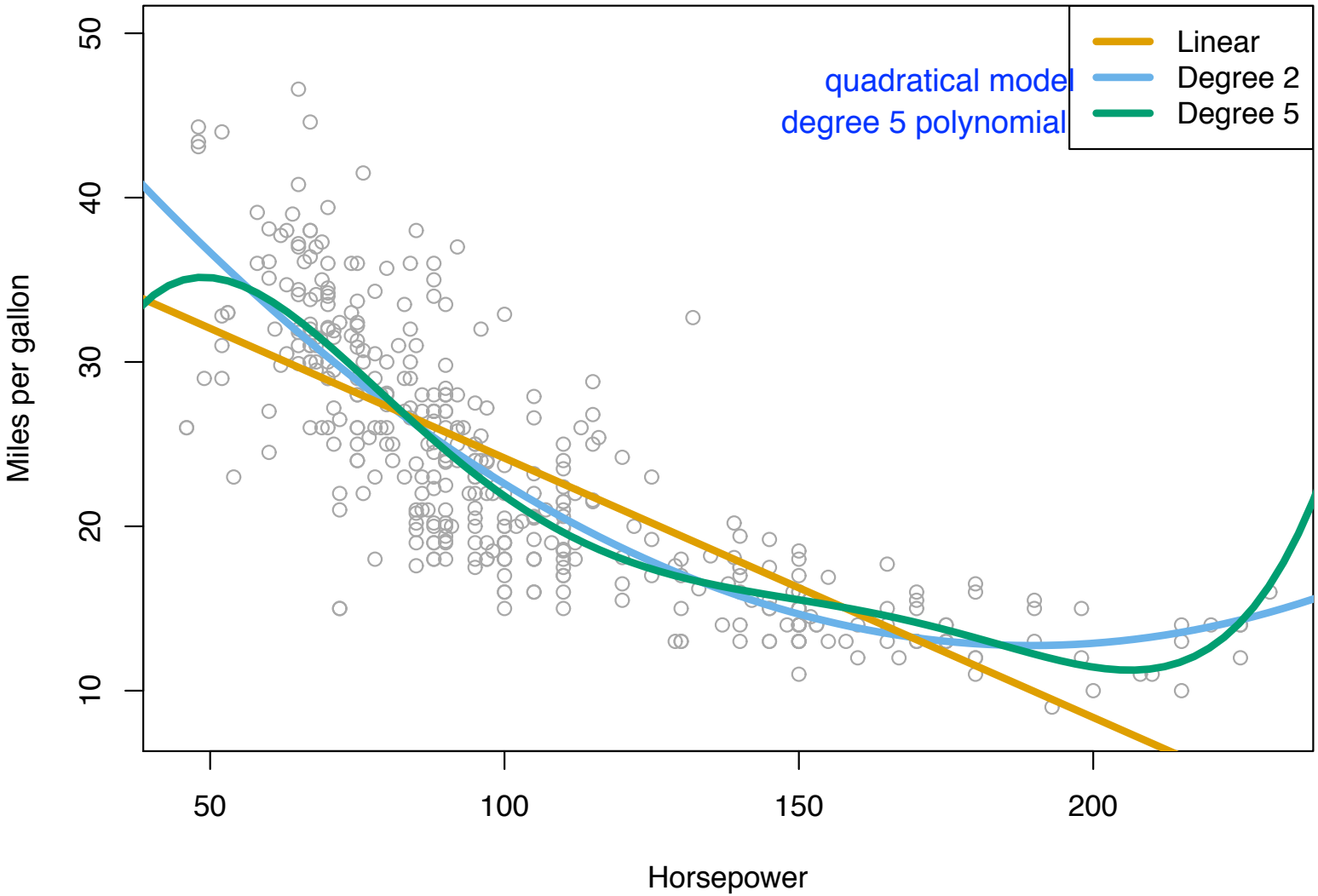


Credit data; Left: no interaction between `income` and `student`. Right: with an interaction term between `income` and `student`.

The other modification of the linear model is what if we want to include nonlinear effects?

Non-linear effects of predictors

polynomial regression on `Auto` data



The figure suggests that

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

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

So that's a very easy way of allowing for nonlinearities in a variable, and still use linear regression. We still call it a linear model, because it's actually linear in the coefficients. But as a function of the variables, it's become nonlinear.

What we did not cover

Outliers
Non-constant variance of error terms
High leverage points
Collinearity

See text Section 3.33

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; nearest neighbor methods.
- *Interactions:* Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- *Regularized fitting:* Ridge regression and lasso