Linear Regression

Potential Problems

Linear regression

- When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

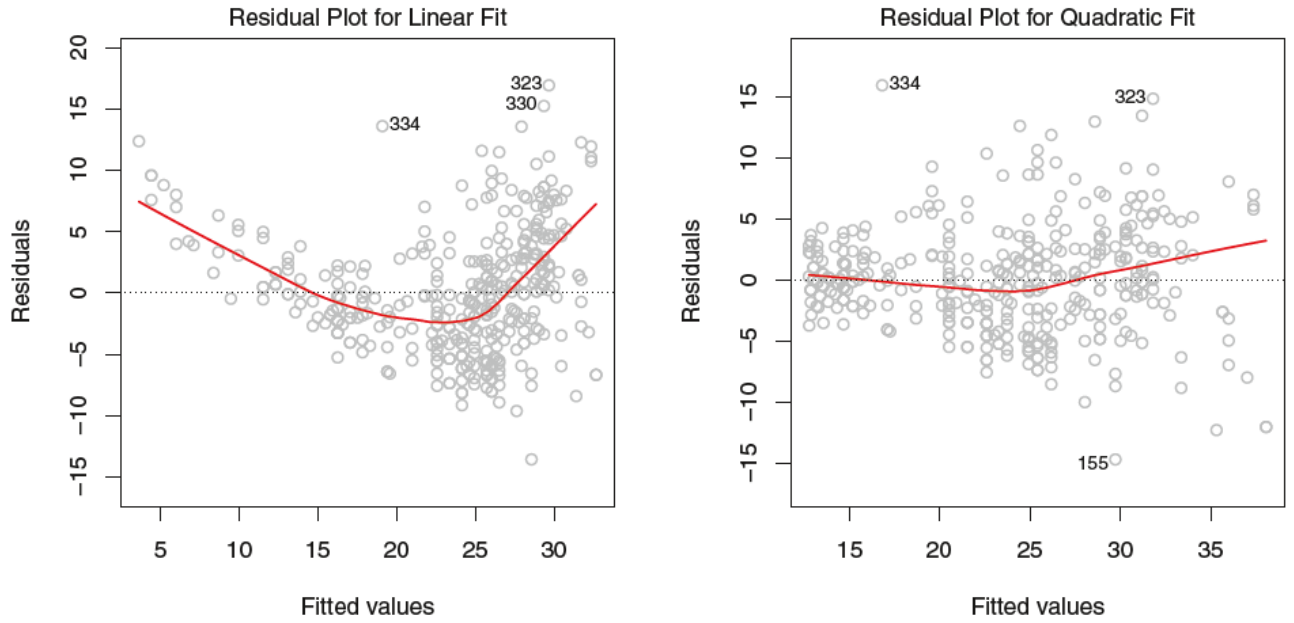
1. *Non-linearity of the response-predictor relationships.*
2. *Correlation of error terms.*
3. *Non-constant variance of error terms.*
4. *Outliers.*
5. *High-leverage points.*
6. *Collinearity.*

1. Non-linearity of the Data

- Consider the advertising data shown on the next slide.
- 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.

Residual plots

- *Residual plots* are a useful graphical tool for identifying non-linearity.
- Given a simple linear regression model, we can plot the residuals,
- $\epsilon_i = y_i - \hat{y}_i$ versus the predictor x_i.



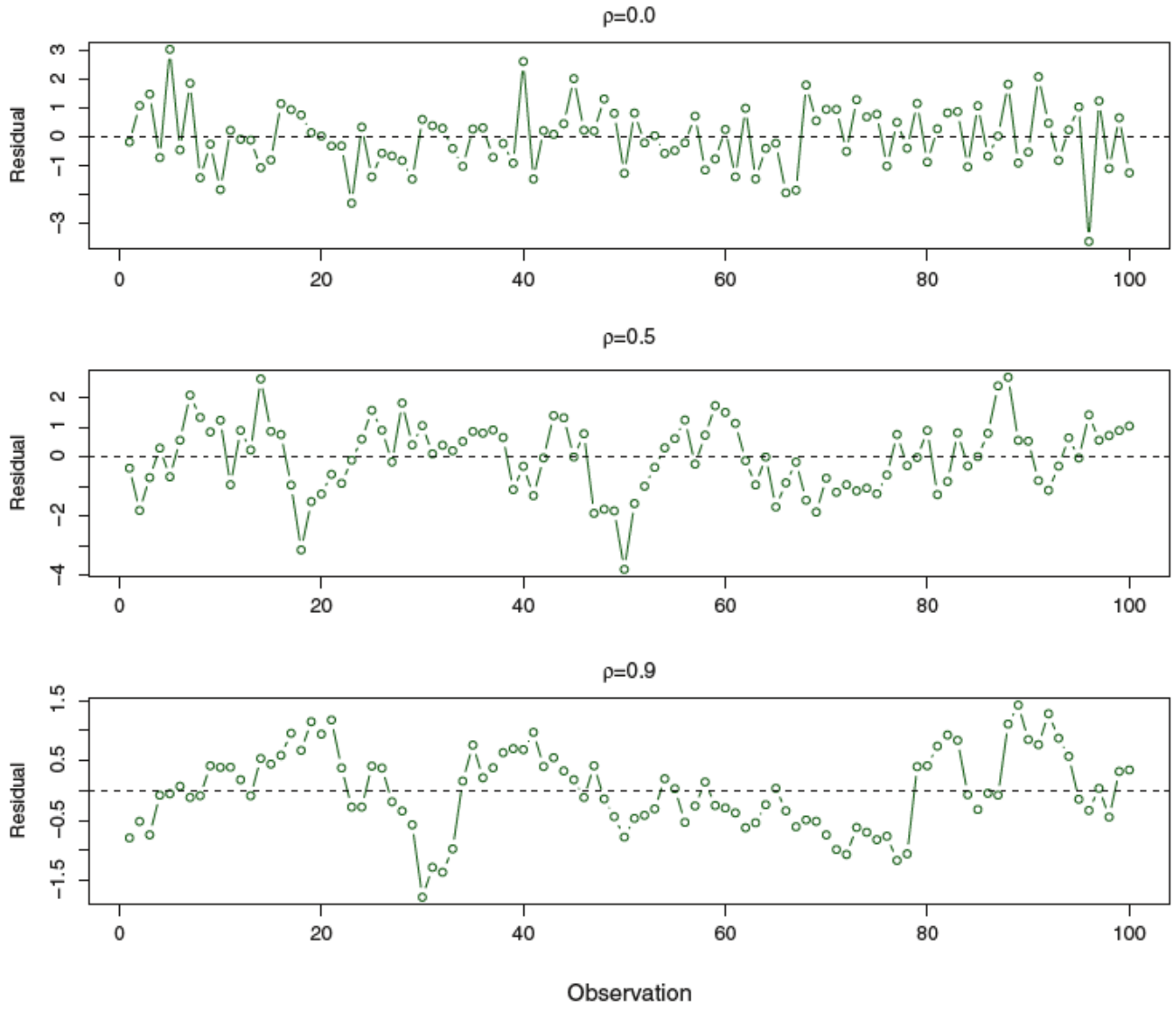
- *Plots of residuals versus predicted (or fitted) values for the* Auto *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* mpg *on* horsepower*. A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* mpg *on* horsepower *and* horsepower^2*. There is little pattern in the residuals.*

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.
- 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 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.
- For example, a 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter.
- In addition, p-values associated with the model will be lower than they should be.

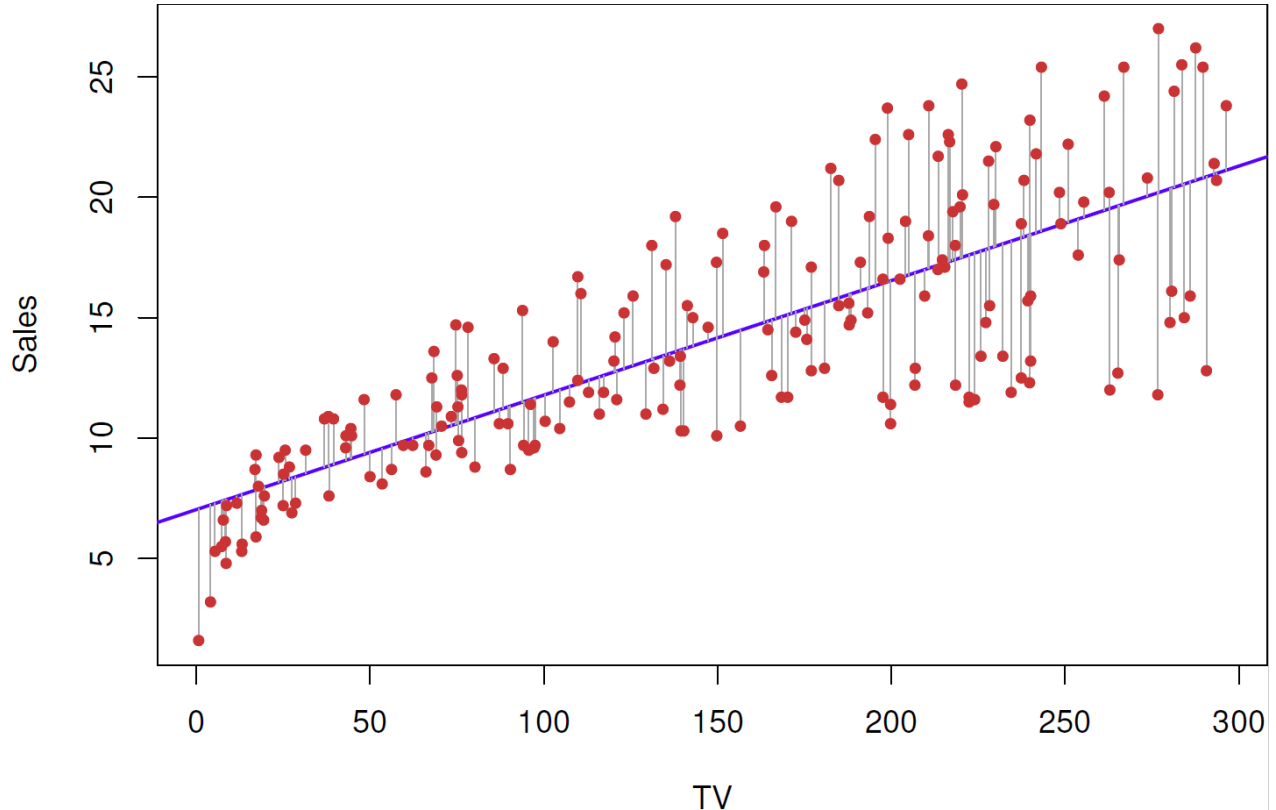
Correlation of Error Terms

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



Example: advertising data

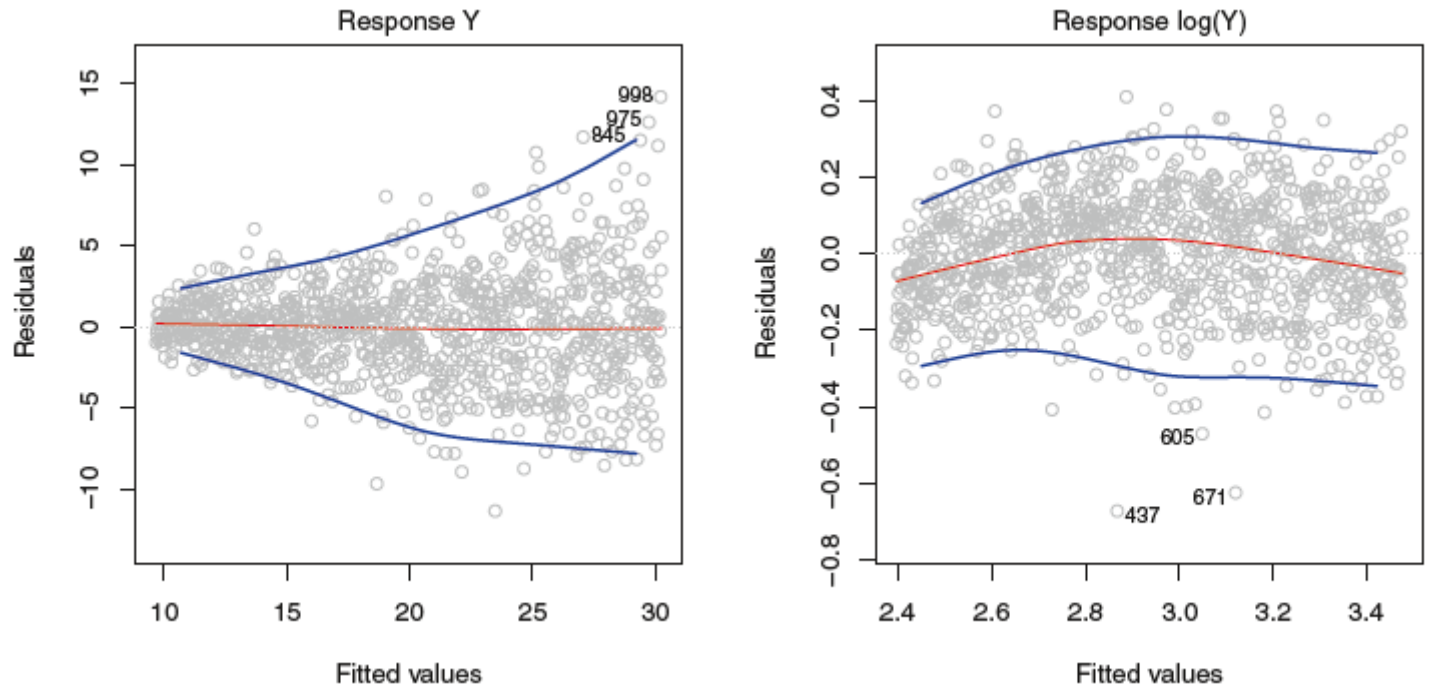
- The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot



3. Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{Var}(\epsilon_i) = \sigma^2$.
- The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption.
- 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.
- When faced with this problem, one possible solution is to transform the response *Y* using a concave function such as log *Y* or $\sqrt{Y}$
- Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.

Non-constant Variance of Error Terms



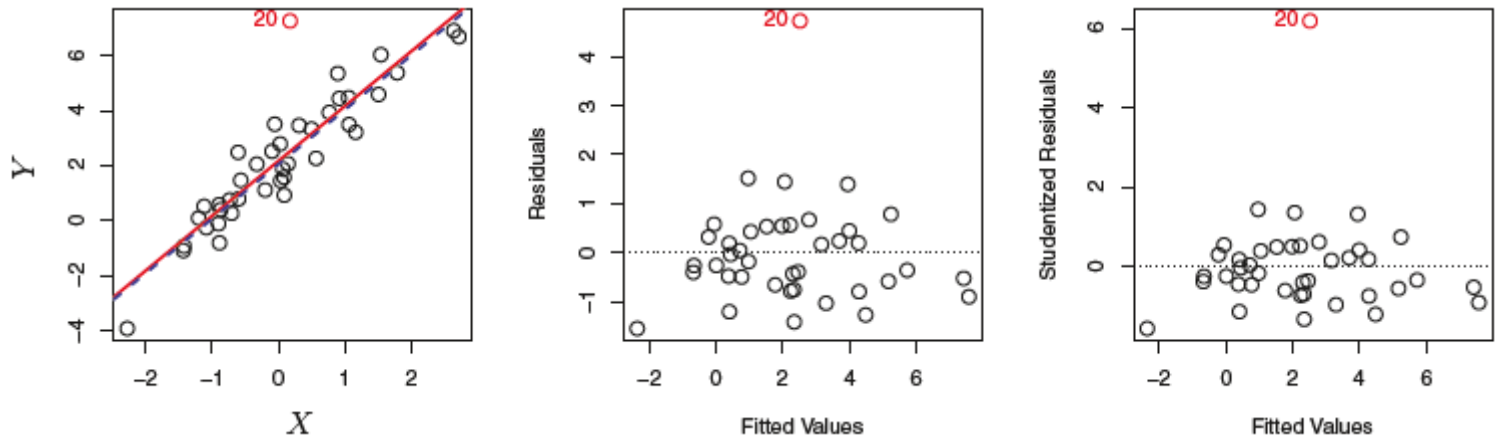
- *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

Non-constant Variance of Error Terms

- Sometimes we have a good idea of the variance of each response. For example, the *i*th response could be an average of n_i raw observations.
- If each of these raw observations is uncorrelated with variance σ^2 $\sigma 2$, then their average has variance $\sigma_i^2 = \sigma^2/n_i$.
- In this case a simple remedy is to fit our model by *weighted least squares*, with weights proportional to the inverse weighted variances—i.e. $w_i = n_i$ in this case.

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.



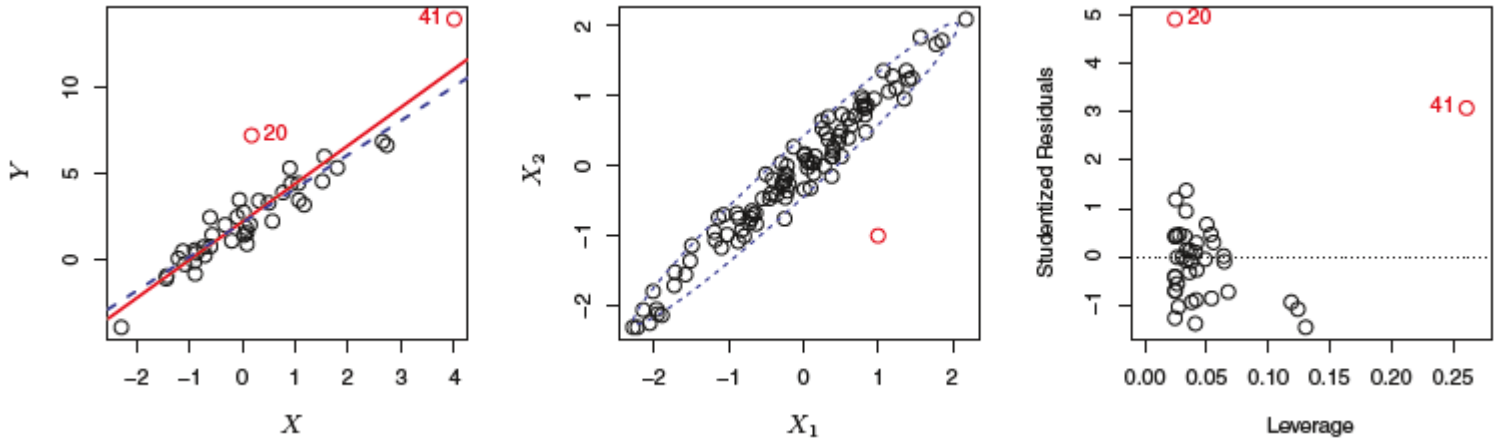
- Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of* 6*; typically we expect values between* −3 *and* 3.

Outliers

- It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit.
- Even if an outlier does not have much effect on the least squares fit, it can cause other problems.
- 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 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.
- Similarly, inclusion of the outlier causes the R^2 to decline from 0.892 to 0.805.

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* high leverage have an unusual value for x.



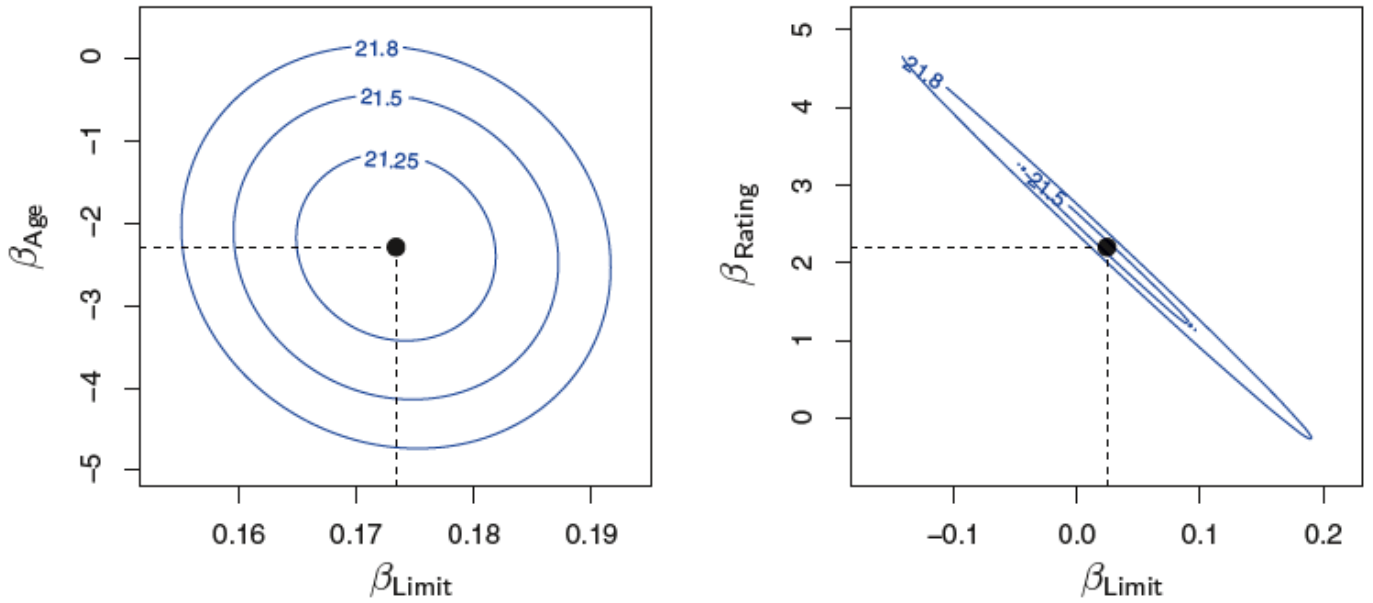
Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its* X_1 *value or its* X_2 *value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

High Leverage Points

- In a simple linear regression, high leverage observations are fairly easy to identify, since we can simply look for observations for which the predictorvalue is outside of the normal range of the observations.
- In a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors. An example is shown in the center panel of the previous slide.
- 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. 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 statistic h_i is always between 1/*n* and 1, and the average leverage for all the observations is always equal to (*p*+1)/*n*.

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 in the contour plots below using the Credit data set.
- In the left-hand panel of Figure, the two predictors limit and age appear to have no obvious relationship. We see that the true limit coefficient is almost certainly somewhere between 0.15 and 0.20.
- 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*. Now 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. the scale for the limit coefficient now runs from roughly −0.2 to 0.2; this is an eight-fold increase



Multiple Regression Collinearity

- *The results for two multiple regression models involving the* Credit *data set are shown below. Model 1 is a regression of* balance *on* age *and* limit*, and Model 2 a regression of* balance *on* rating *and* limit*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

		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

- Collinearity results in a decline in the *t*-statistic. As a result, in the presence of collinearity, we may fail to reject *H*0 : $\beta_j = 0$. This means that the *power* of the hypothesis test—the probability of correctly power detecting a *non-zero* coefficient—is reduced by collinearity.

Multiple Regression Collinearity

- A simple way to detect collinearity is to look at the correlation matrix of the predictors. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.
- Unfortunately, not all collinearity problems can be detected by inspection of the correlation matrix:
- A better way to assess multi-collinearity is to compute the *variance inflation factor* (VIF).
- The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.
- The VIF for each variable can be computed 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.
- If $R^2_{x_j|x_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

Collinearity

- When faced with the problem of collinearity, there are two simple solutions.
- The first is to drop one of the problematic variables from the regression. This can usually be done without much compromise to the regression fit, since the presence of collinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables.
- For instance, if we regress balance onto age and limit without the rating predictor, then the resulting VIF values are close to the minimum possible value of 1, and the R2 drops from 0.754 to 0.75.
- The second solution is to combine the collinear variables together into a single predictor. For instance, we might take the average of standardized versions of limit and rating in order to create a new variable that measures *credit worthiness*.

Estimation and Prediction for Multiple Regression

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

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

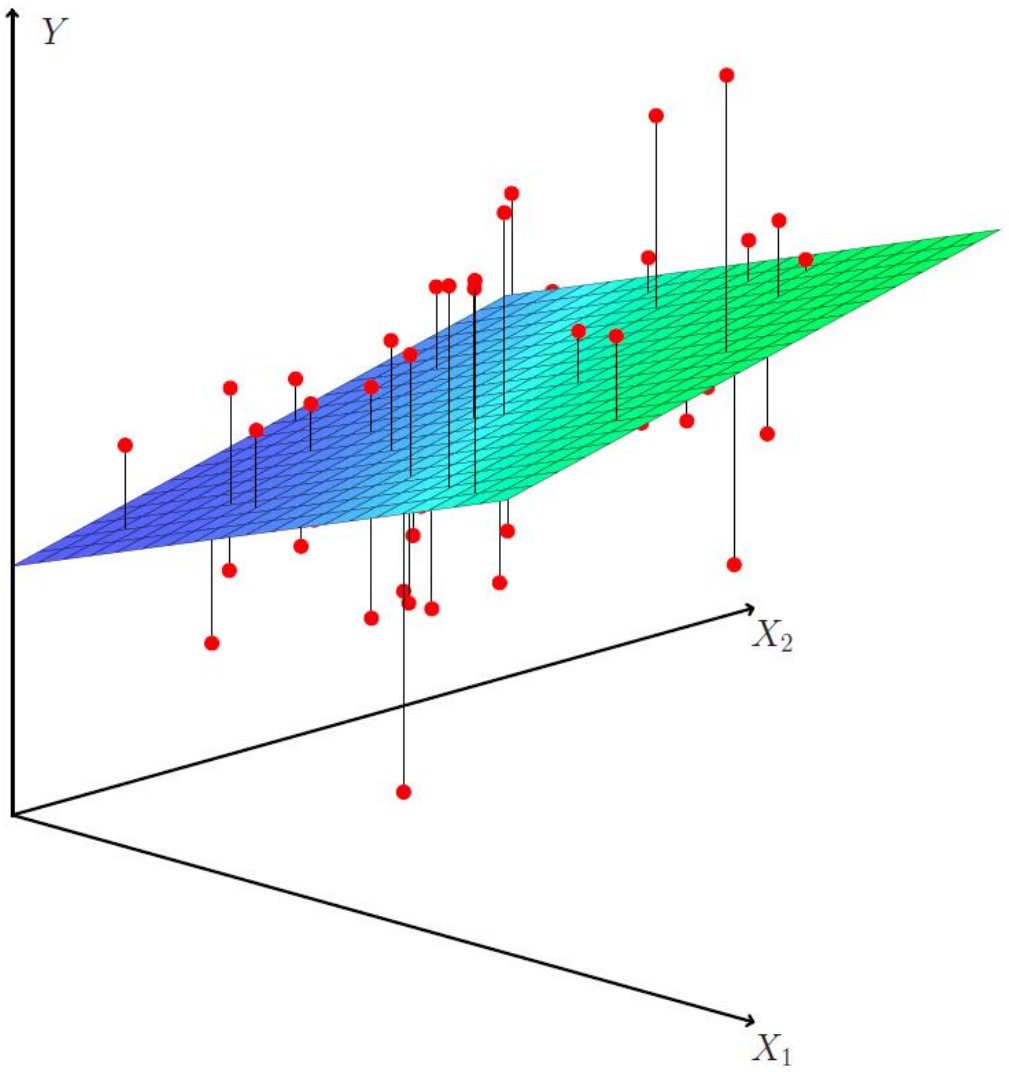
We estimate $\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p$ as the values that minimize the sum of squared residuals

- 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 $\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

Linear approximation to advertising data



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

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

Some important questions

1. Is at least one of the predictors X_1, X_2 , … 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

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 there 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

- 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.

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:

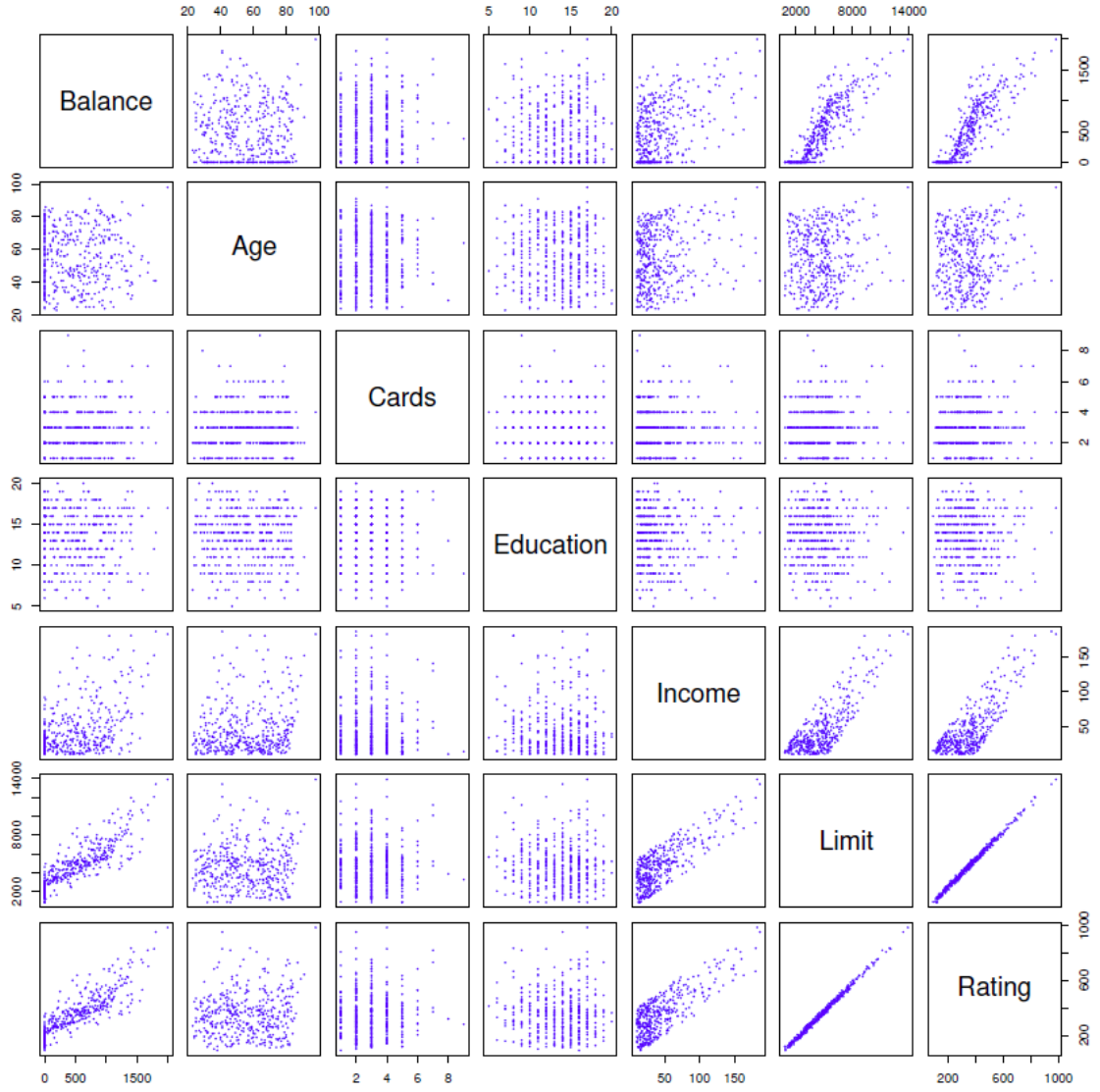
Mallow’s C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 and Cross-validation (CV).

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:

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

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

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

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

- 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*.

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

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.

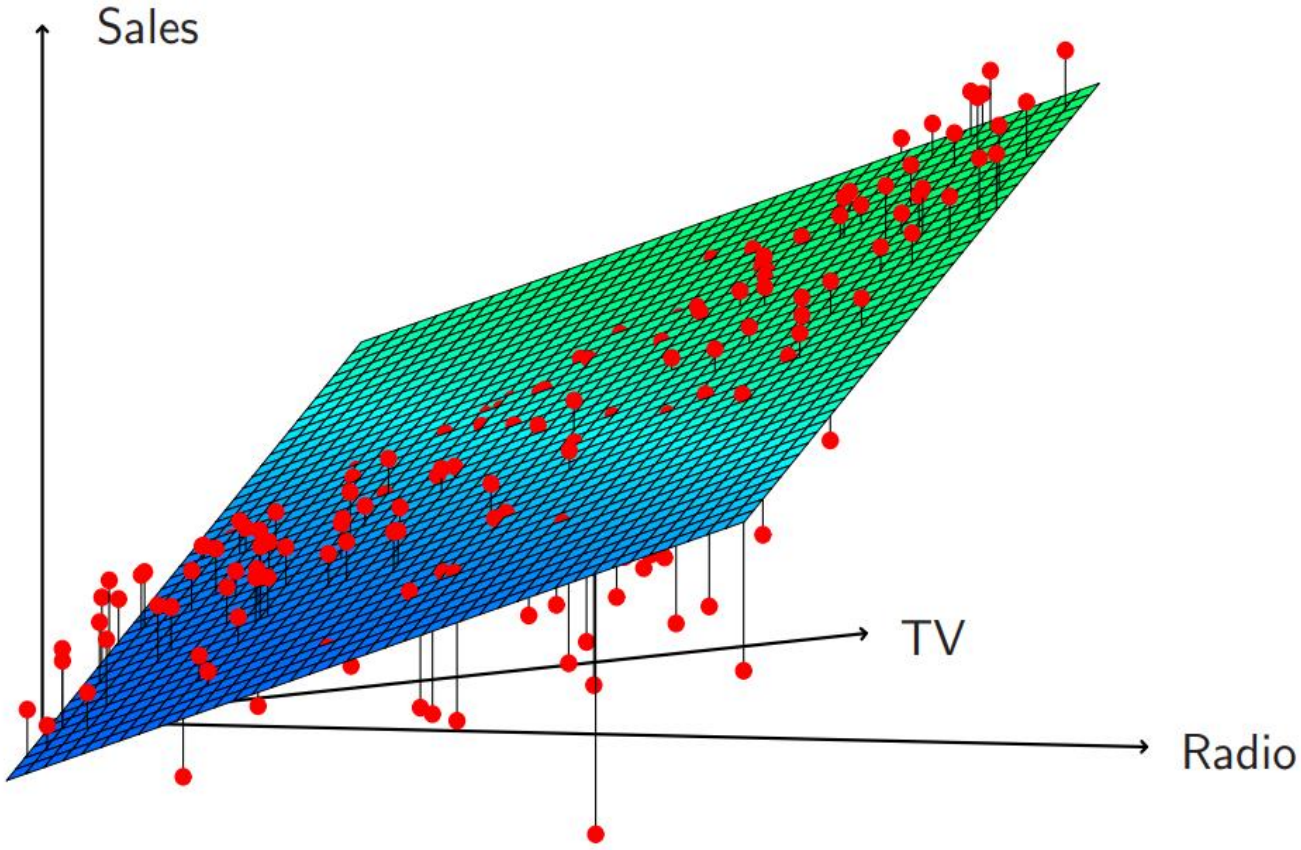
For example, the linear model

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

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.

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



Modeling interactions — Advertising data

Model takes the form

$$\widehat{\text{sales}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \epsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Results:

	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

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.
- 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

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

  $(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$ 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.

- 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

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

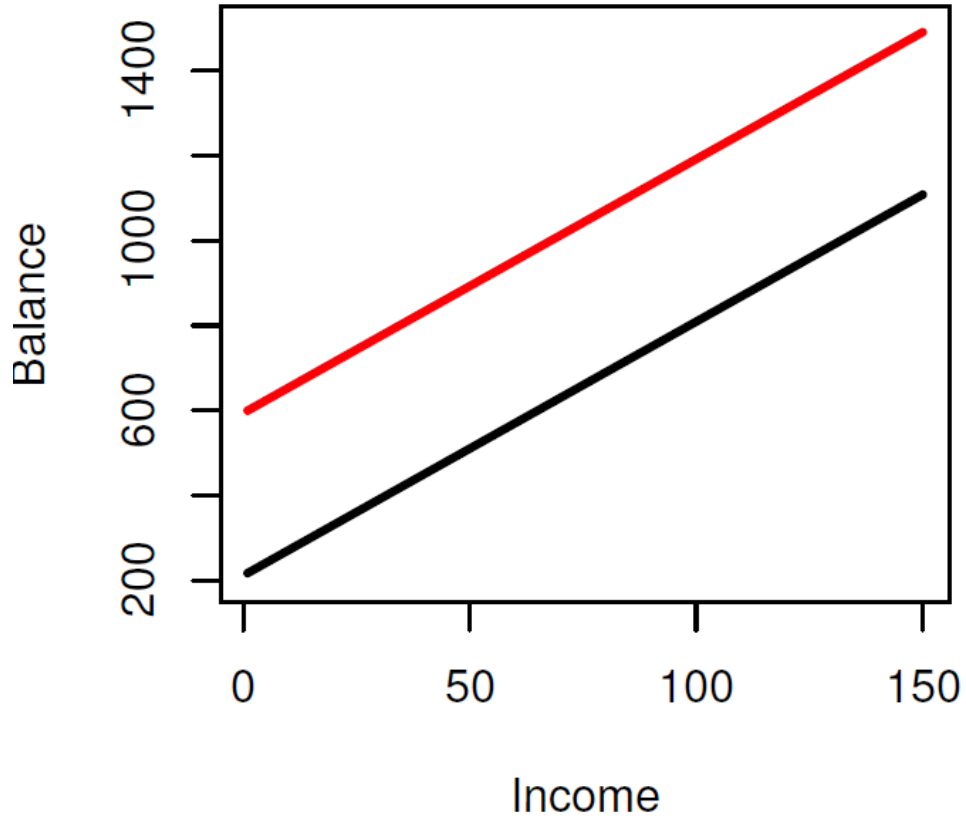
$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases}$$

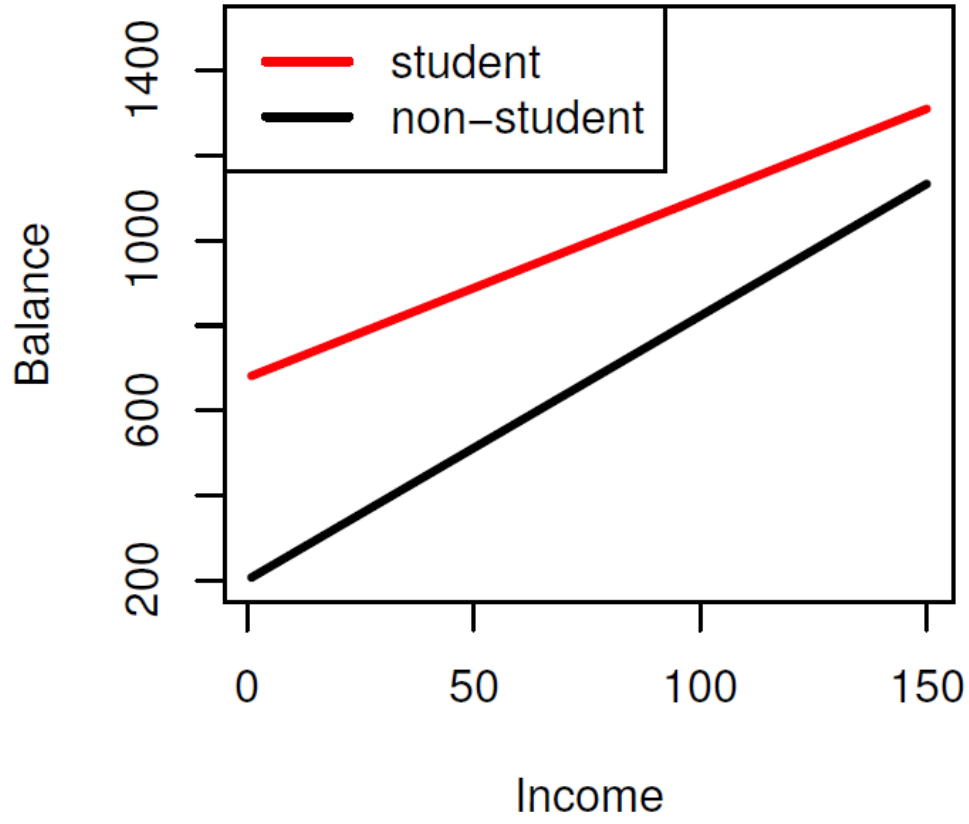
Interactions between qualitative and quantitative variables

With interactions, it takes the form

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

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





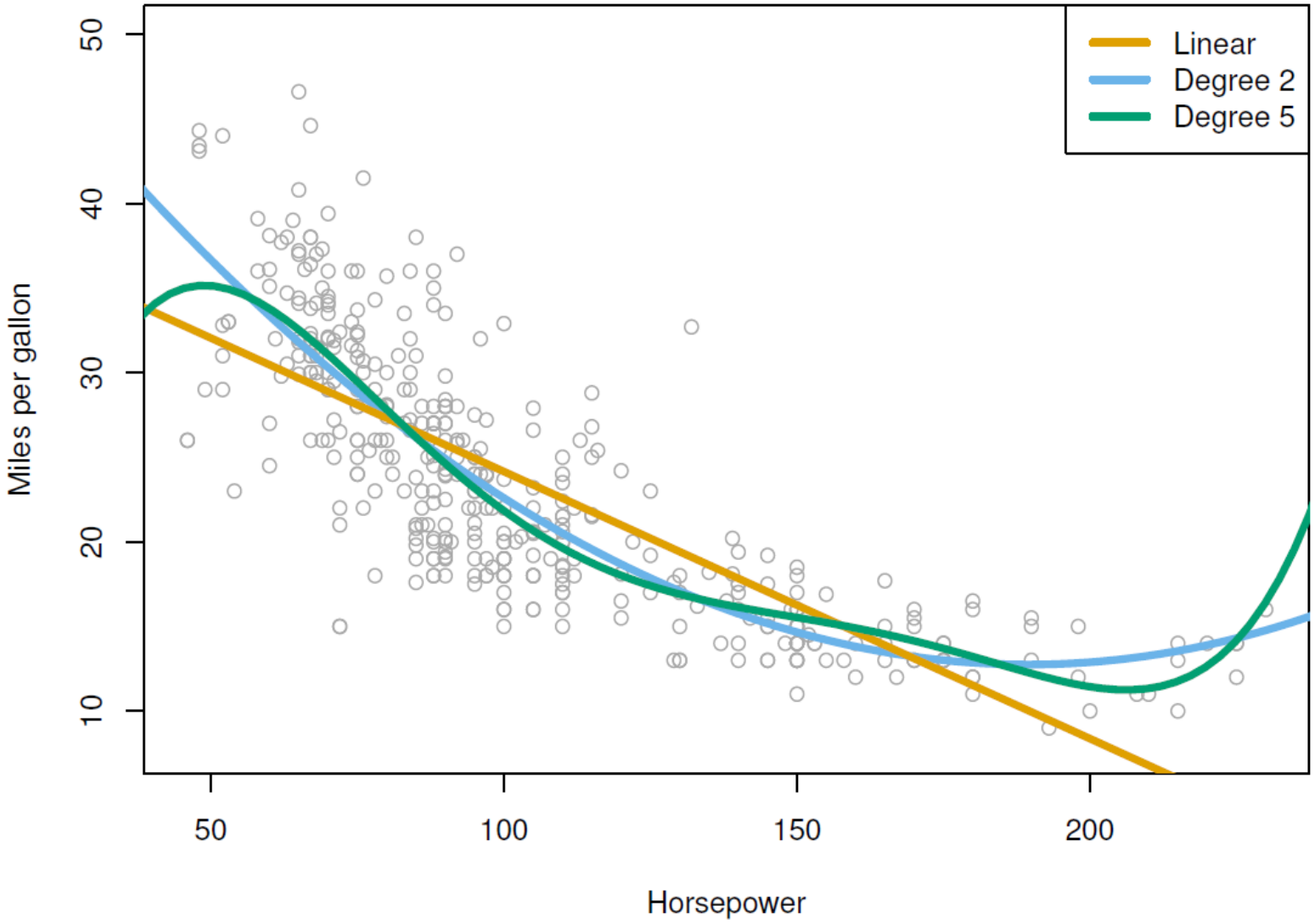
Credit data;

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



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