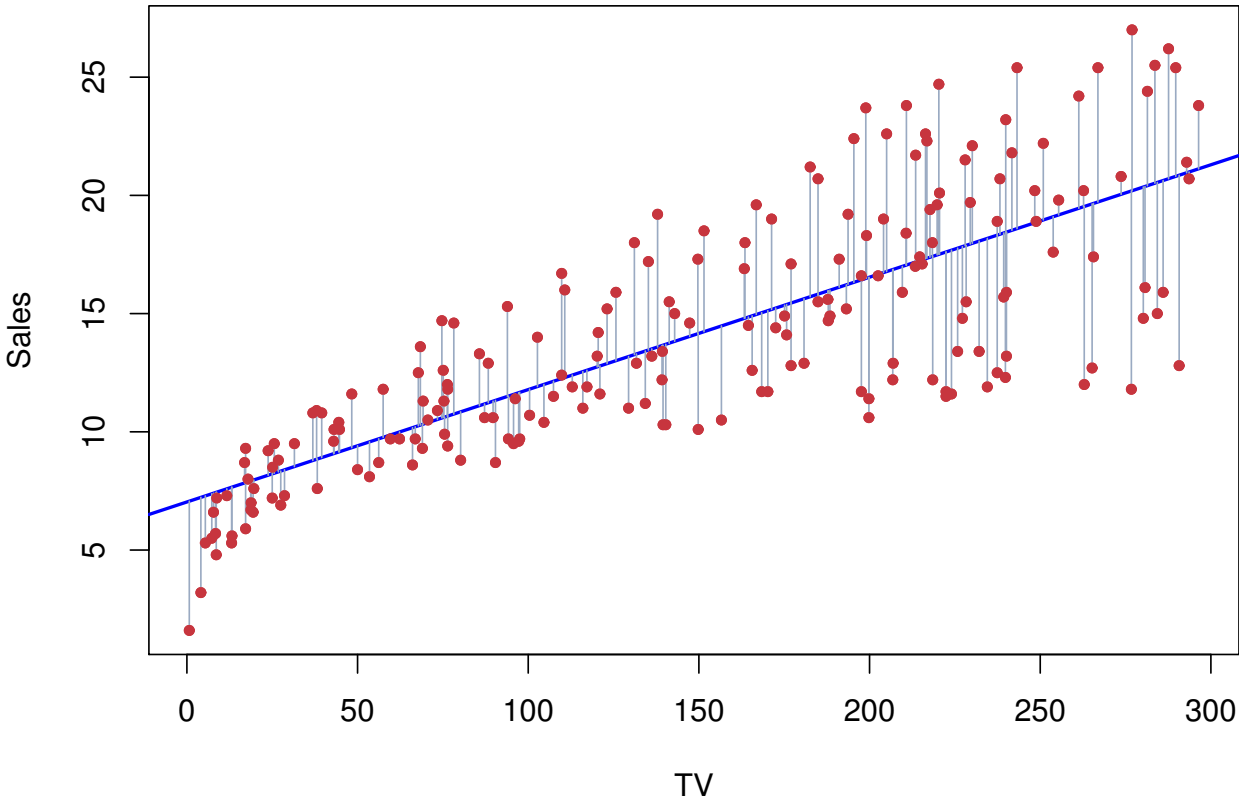
Chapter 03 – Linear Regression

Slides by Zia Khan

Simple Linear Regression



Residuals
Thin blue lines.

$$Y \approx \beta_0 + \beta_1 X.$$

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

Residual Sum of Squares (RSS)

$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$ Training data

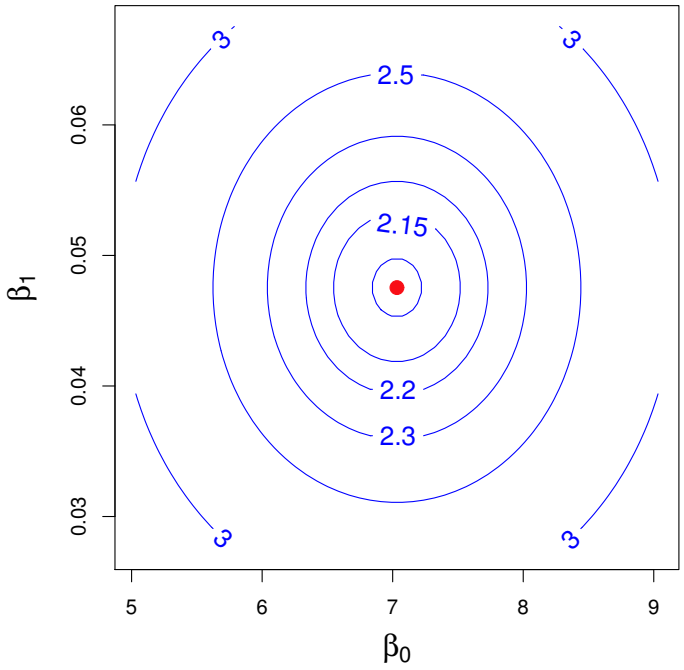
$$e_i = y_i - \hat{y}_i$$ Residual – difference between ith observed response value and ith predicted value from linear model

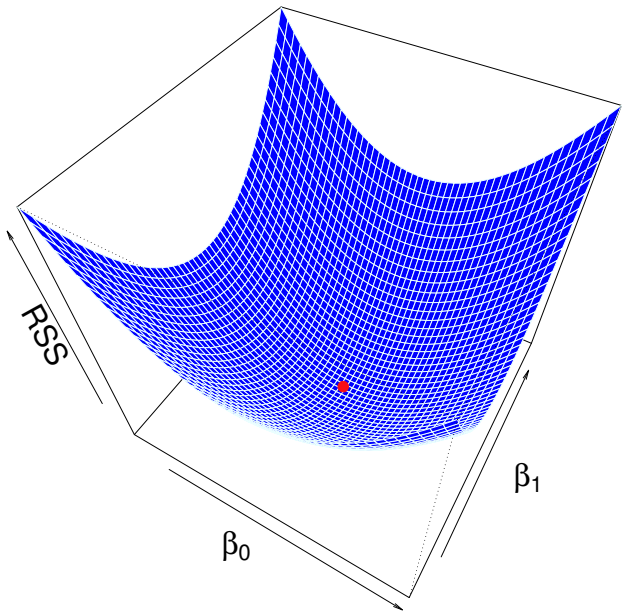
$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

Residual sum of squares.

Least squares fit choses betas that minimize RSS.

RSS



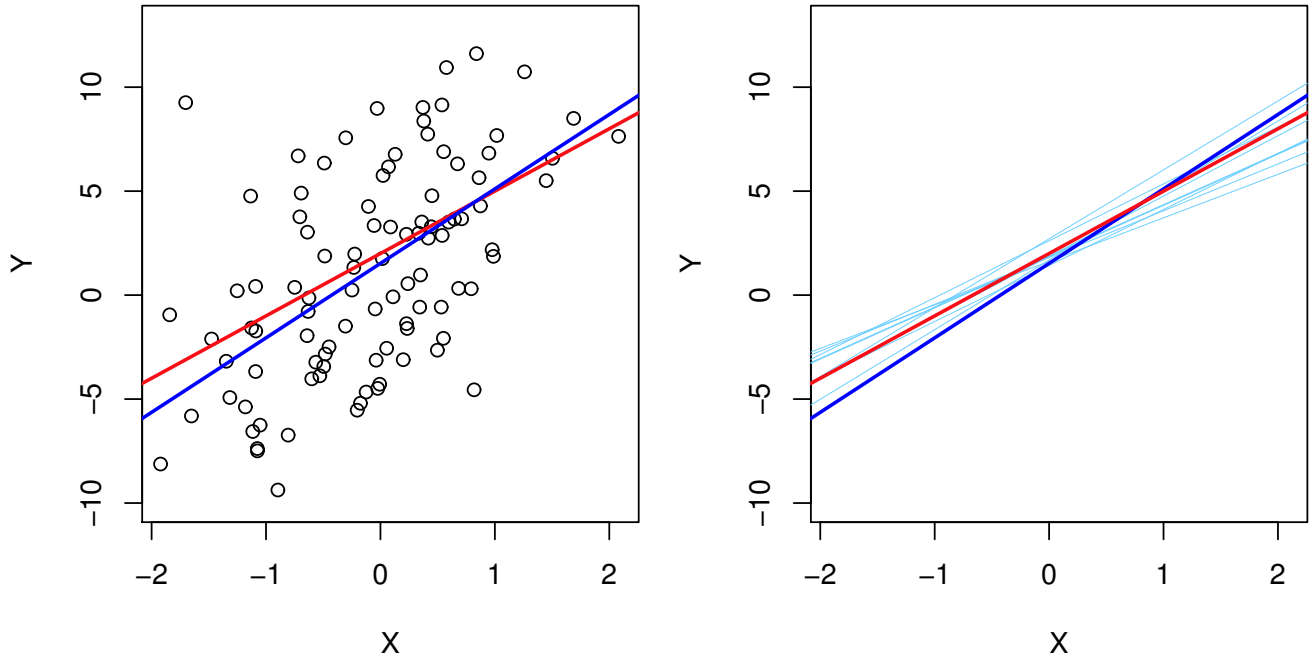


$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

Minimize RSS (derived using some calc).

Population Regression Line



Population regression line is unobserved true relationship.
Blue is the least square regression line for a sample.
Light blue lines are least squares regression lines for many samples.
If we average these regression lines over a large number of data sets, the result approaches population regression line.
Least squares estimate of parameters is **unbiased**.

Standard Error

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

Standard error of the mean.
Average amount estimate of mean differs from actual mean.
Shrinks with larger n.

For simple linear regression:

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

Confidence Intervals and Hypothesis Testing

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$ 95% confidence interval

Null hypothesis:

H_0 : There is no relationship between X and Y $\qquad H_0 : \beta_1 = 0$

Alternative hypothesis:

H_a : There is some relationship between X and Y. $\qquad H_a : \beta_1 \neq 0,$

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

T-statistic, t distribution with n-2 degrees of freedom

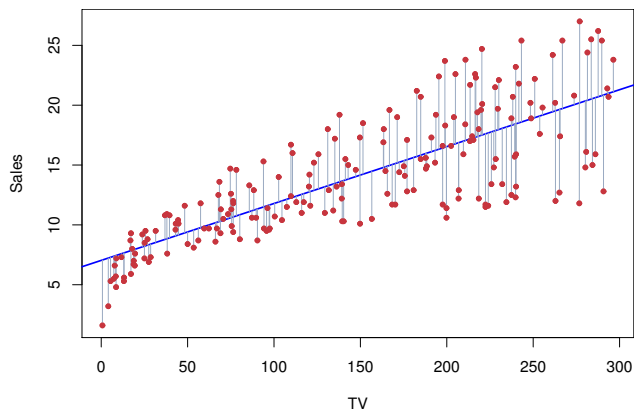
Probability of observing a beta-hat not equal to 0.

p-value and rejecting the null hypothesis

p-value indicates how unlikely it is to observe a beta-hat not equal to zero by chance.

If p-value is small enough, we can reject the null hypothesis and say significant relationship exists between X and Y.

	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



$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$.

TV advertising is significantly associated with sales.

(intercept) in the absence of TV expenditure sales is significantly non-zero.

Residual Standard Error (RSE) and R^2

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

Standard deviation of linear regression error.

Measures lack of fit of linear regression.

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

Proportion of variance explained.

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

Total sum of squares, measures variability of response.

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Measures remaining variability after linear model is fit.

Multiple Linear Regression

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

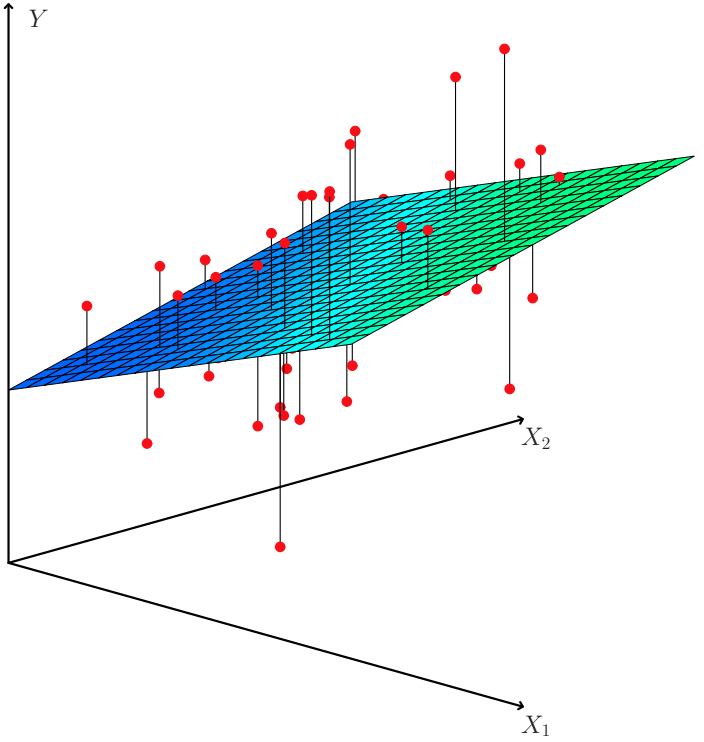
Multiple predictors in regression.
Adjust for correlation among predictors.

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

Minimize RSS To Estimate Regression Coefficients

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

Fits a least squares plane or hyperplane to data.



Is there a relationship between response and predictors?

Null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Alternative hypothesis:

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

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

F-statistic

If F-statistic is > 1 then more evidence against the null.

F-statistic for comparing models

Null hypothesis for p-q predictors:

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

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}.$$

Does adding q predictors to the model have a significant effect.

Do these q new predictors have a significant effect, control for the remaining (p – q) predictors?

Residual Standard Error for Multiple Linear Regression

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}},$$

Models with more variables can have higher RSE if increase in RSS is small relative to p (number of predictors).

Measures model fit to data.

Qualitative Predictors (or Categorical Predictors) with 2 Levels

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

Dummy variable.

Y here is credit card balance.

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

Beta0 = the average credit card balance for males
Beta0 + Beta1 = average credit card balance for females
Beta1 = average difference between credit card balances for males and females

Qualitative Predictor with 2 Levels: Alternate Coding Scheme

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

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

beta0 = overall average credit card balance
beta1 = amount females are above average and males below average

Different coding scheme gives predictors a different interpretation.

More than 2 Levels for Qualitative (or Categorical) Predictors

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

e.g.
ethnicity = { Asian, Caucasian, African American }

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

y is credit card balance again.

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

beta0 = average credit card balance for African Americans
beta1 = difference between African American and Asian categories
beta2 = difference between African American and Caucasian categories

Coding schemes allow certain contrasts and change interpretation of the betas.

Interactions in Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

Standard linear model is additive and linear.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

Product adds an interaction term.

Re-write as:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

Adjusting X_2 will change the impact of X_1 on Y.

Interaction: Example for Quantitative Predictors

$$\begin{aligned}\texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} + 1.4 \times (\texttt{lines} \times \texttt{workers}) \\ &= 1.2 + (3.4 + 1.4 \times \texttt{workers}) \times \texttt{lines} + 0.22 \times \texttt{workers}.\end{aligned}$$

Effect of adding additional assembly lines will increase with more workers.

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

beta3 is increase in effectiveness of TV advertising for a unit increase in radio advertising and vice-versa

Interaction: Between Quantitative and Qualitative Variable

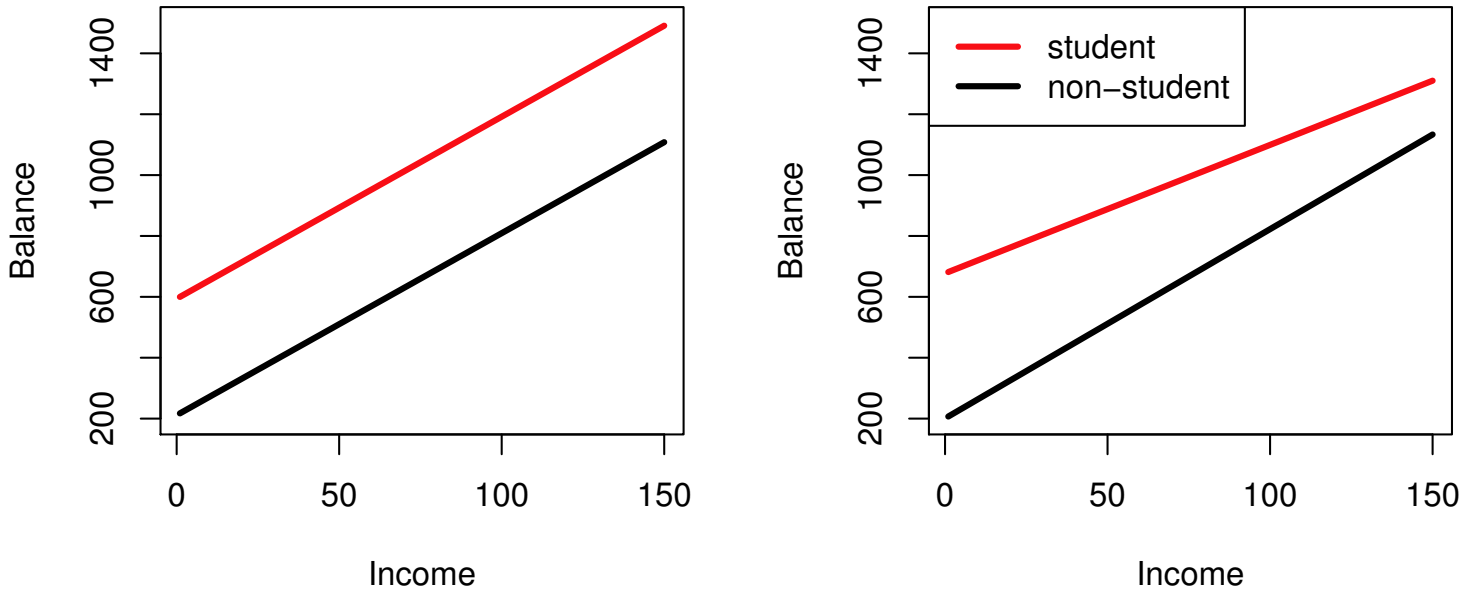
$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if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

No interaction: Common slope between students and non-students relating income to credit card balance. Yet, intercept is different.

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

Interaction between income and student status allows different slopes and intercepts.

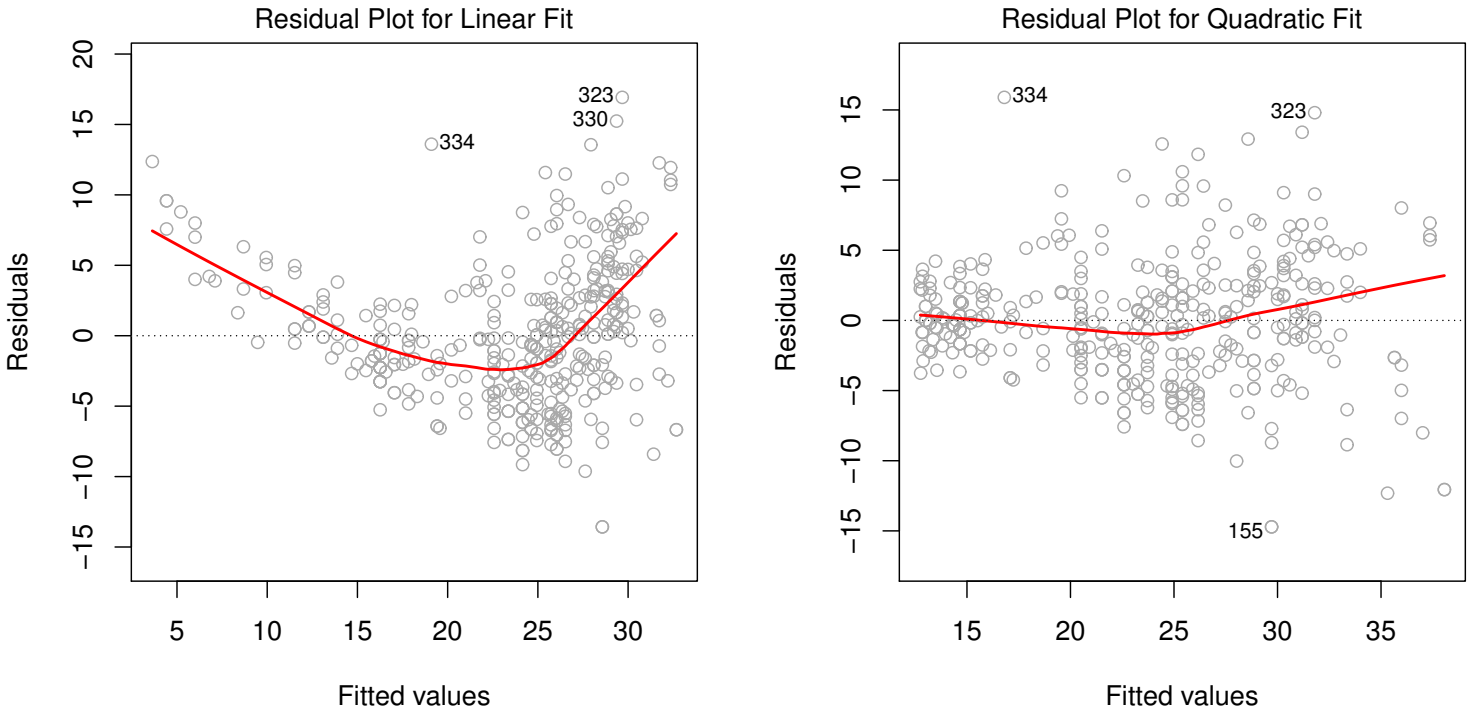
Interaction Between Income and Student Status



Problems with Linear Regression

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.*

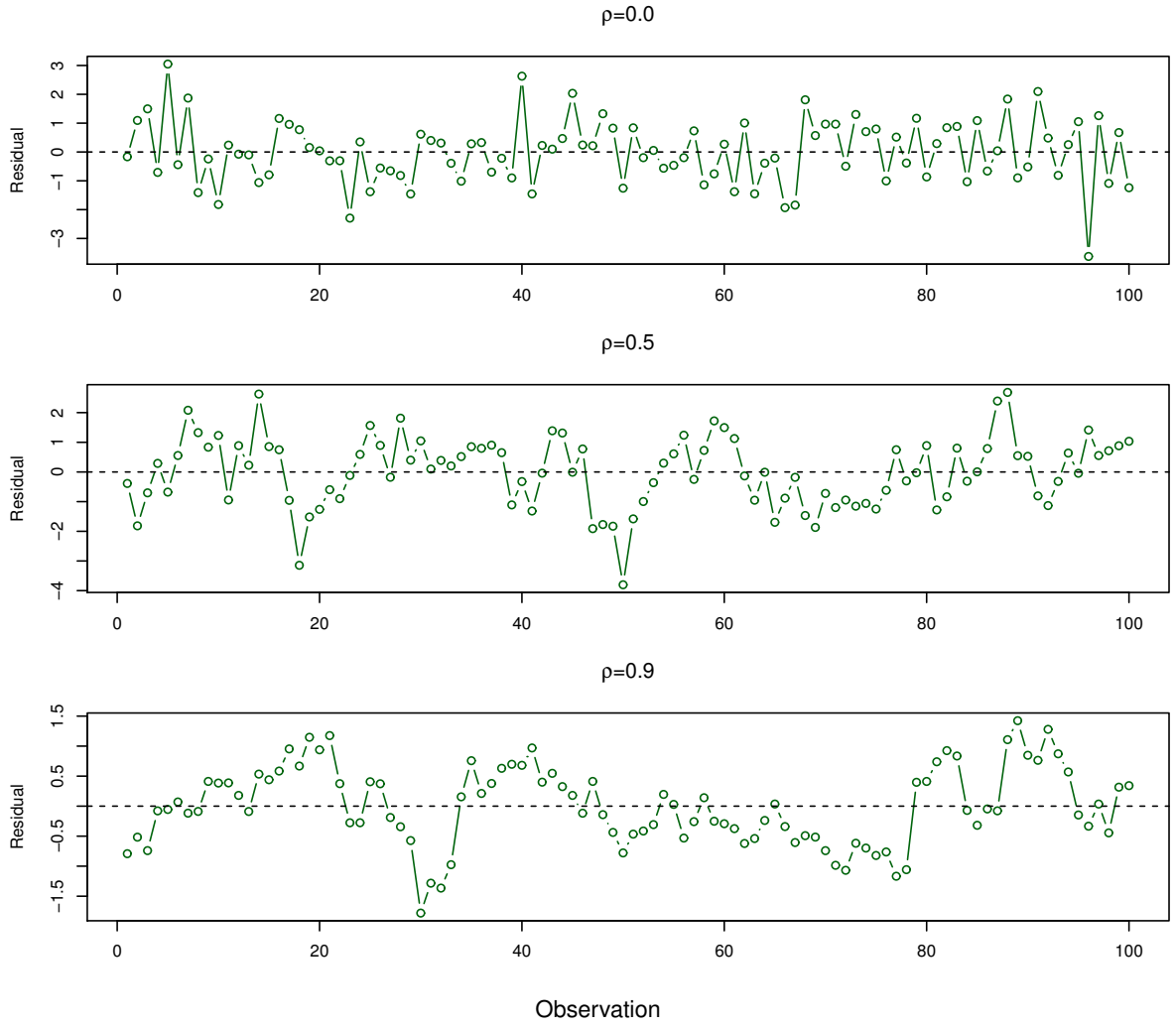
1. Residual Plots and Nonlinearity



Residuals in left plot reveal nonlinearity.

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

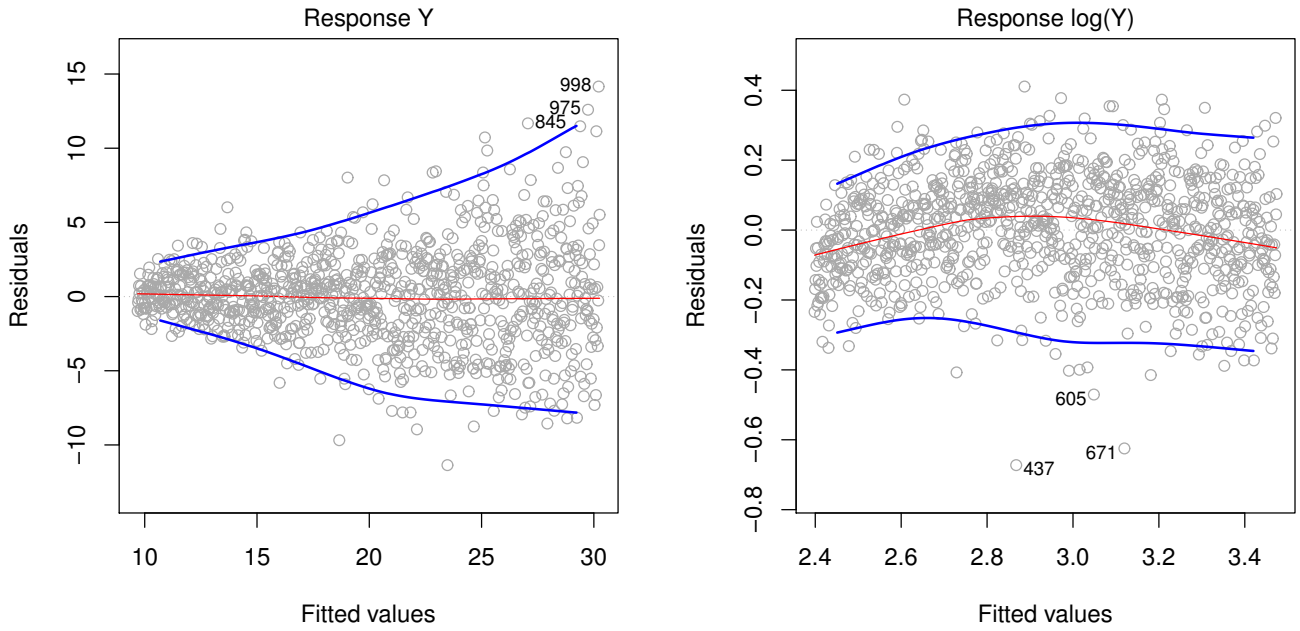
2. Correlated Errors



Leads to underestimated standard errors.

Rho here is the correlation between successive points.

3. Non-Constant Variance of Error



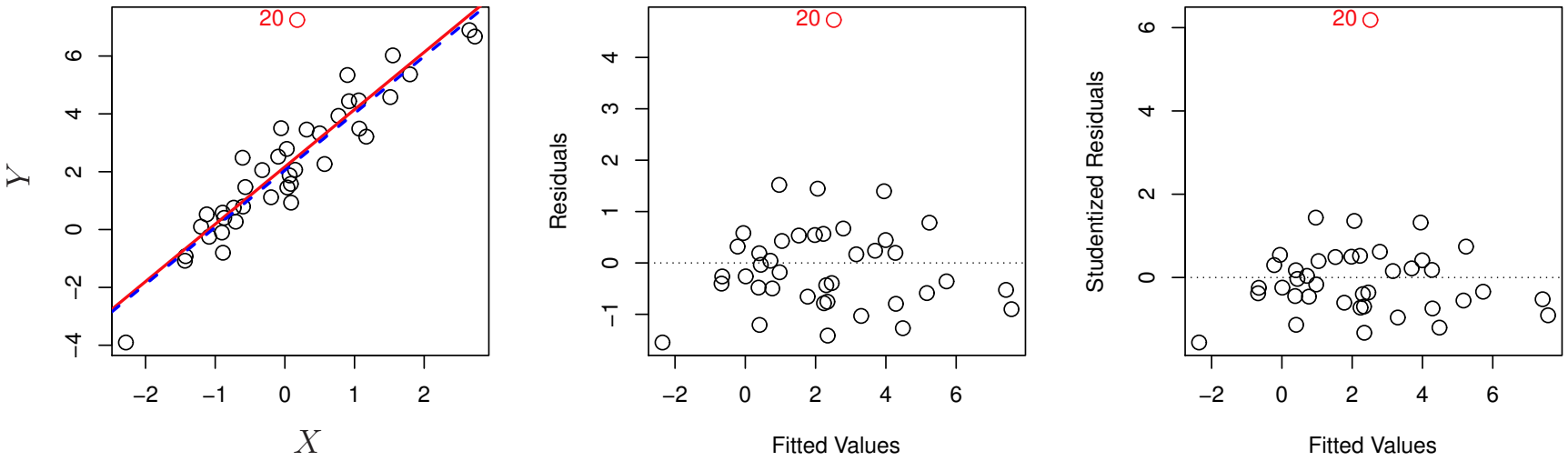
Heteroscedasticity = non-constant variance of error terms

Possible fix. Transform response here log(Y).

Another fix. If you know variance of each observation, fit using a weight. Higher weight for smaller variance.

4. Outliers

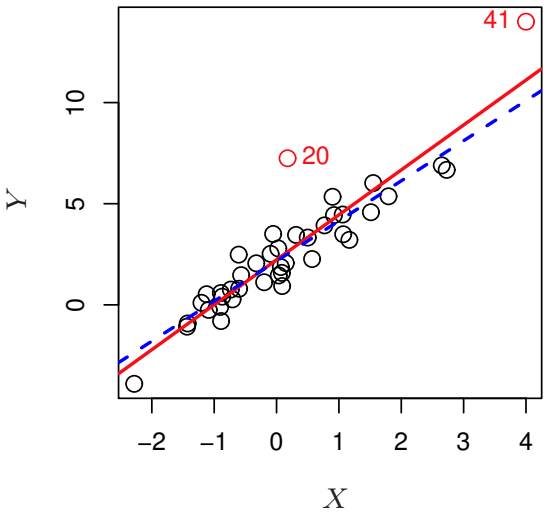
Lead to over estimation of RSE and R^2

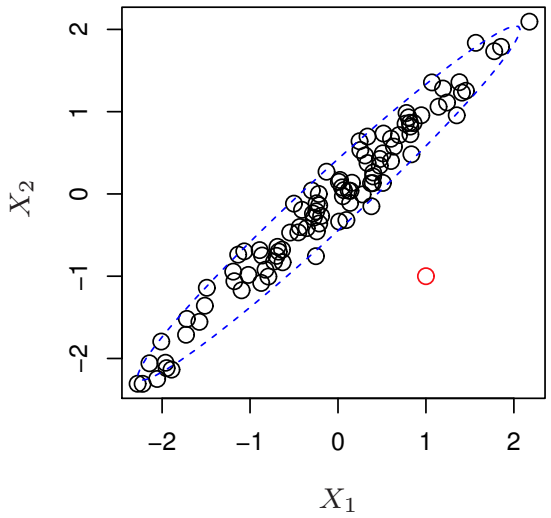


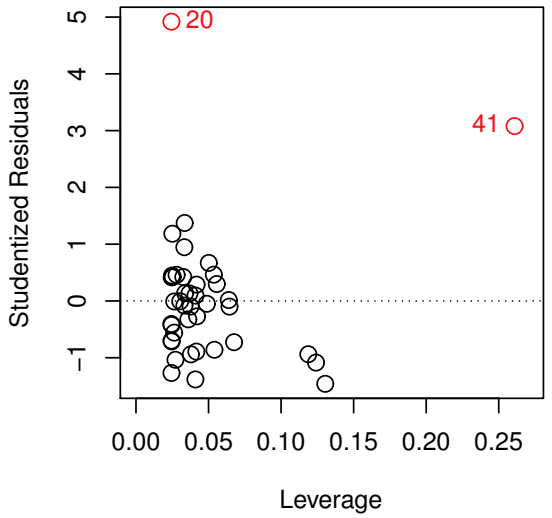
Have a Y value that is well outside what is predicted from linear regression.

Can be identified in residual plots.

5. High Leverage Points



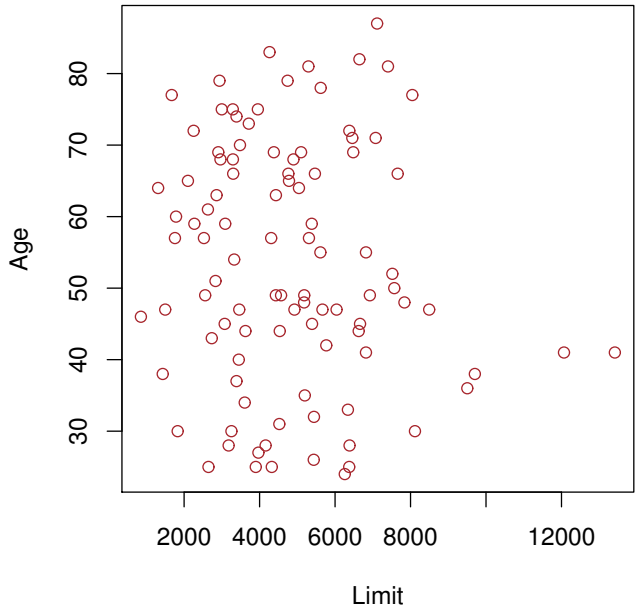


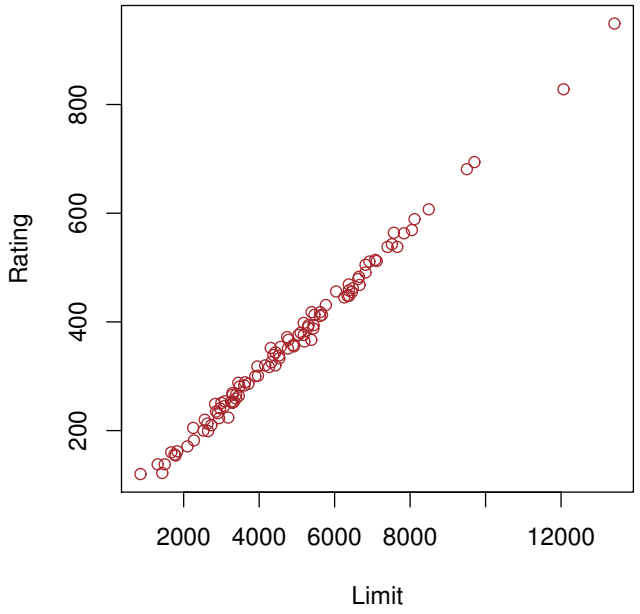


Point has reasonable predicted value, but has an unusual predictor X value.
High leverage point will influence the fit.
More pronounced problem in multiple linear regression.
Can be addressed in part by computing leverage statistics e.g.

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

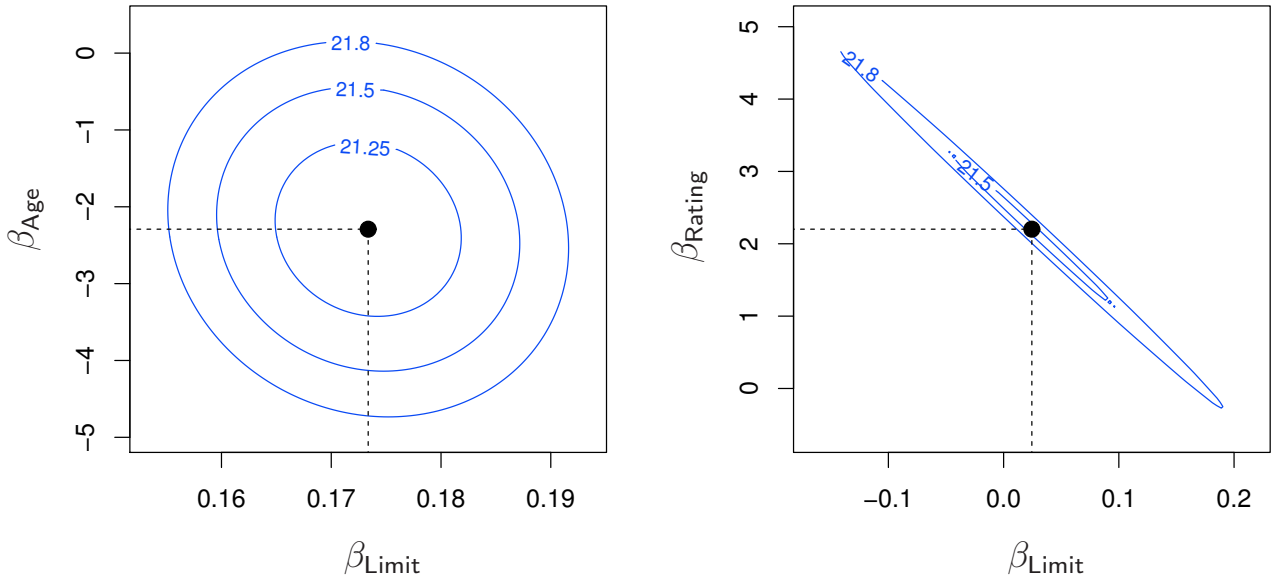
6. Collinear Points





Two or more predictors are closely related.

RSS for as a function of values for betas for collinear predictors



Many betas for which RSS is minimized.
Causes standard errors of betas to be high and you won't detect non-zero betas.

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

Diagnose by removing predictor and computing variance inflation factor.

Regress predictor onto each other predictor.

Linear Regression

- Powerful technique.
- Interpretability is high.
- The first technique you should consider using when addressing data analysis problems.
- Important to use diagnostics to avoid incorrect inferences.