

LECTURE 3: LINEAR REGRESSION

STAT 1361/2360: STATISTICAL LEARNING AND DATA SCIENCE

University of Pittsburgh
Prof. Lucas Mentch



- First supervised learning method: linear regression
 - ▶ **Read the Book (ISLR Ch. 3)**; slides are not meant to be a complete review
- Parametric statistical method
- Regression; we'll assume the response $Y \in \mathbb{R}$, though the method can be used for ordinal categorical responses in some cases (more on this in chapter 4)
- Amongst the oldest yet still most common modeling techniques
 - ▶ Regression and least-squares dating back to early Legendre and Gauss in early 1800s



Recall that in a standard regression framework, we assume

$$Y = f(X) + \epsilon$$

where X is simply shorthand for the predictors (X_1, \dots, X_p) . A **linear model** is defined as one in which f is modeled as a linear function of parameters $\beta = (\beta_0, \dots, \beta_p)$

$$\text{e.g. } Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

and the parameters correspond to coefficients



Note: The *linear* in linear models refers to the linearity in parameters:

$$\text{e.g. } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$$

$$Y = \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 (X_1 e^{X_3}) + \epsilon$$

are all linear models.



Why so popular?

- Why is linear regression so popular?
 - ▶ Explicit and interpretable
 - ▶ Many useful corresponding statistics and inference procedures
 - ▶ Does not require a large sample to fit
 - ▶ Many natural relationships can be well-approximated by a suitable linear model
 - ▶ **Not** because we believe many natural relationships actually *are* linear



SIMPLE LINEAR REGRESSION

Simple Linear Regression

We begin with the simplest case, **simple linear regression**: only a single feature X_1 and we model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- Given data $(x_{11}, y_1), \dots, (x_{1n}, y_n)$, we need to estimate parameters $\beta = (\beta_0, \beta_1)$
 - ▶ β_0 = y -intercept (average y -value when $x = 0$)
 - ▶ β_1 = slope (average change in y with 1 unit change in X_1)
- Given parameter estimates $\hat{\beta}_0, \hat{\beta}_1$, our model estimate becomes

$$\begin{aligned}\hat{y}^* &= \hat{f}(x_1^*) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_1^*\end{aligned}$$



Estimating the Parameters

How do we estimate the parameters β_0 and β_1 ?

- Standard statistical method: define a *loss function* and choose parameter estimates that minimize that loss
- Most common choice is squared error loss, or **Least Squares**:
Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the residual sum of squares (RSS)

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i}))^2$$

Minimizing gives *ordinary least squares* (OLS) estimates:

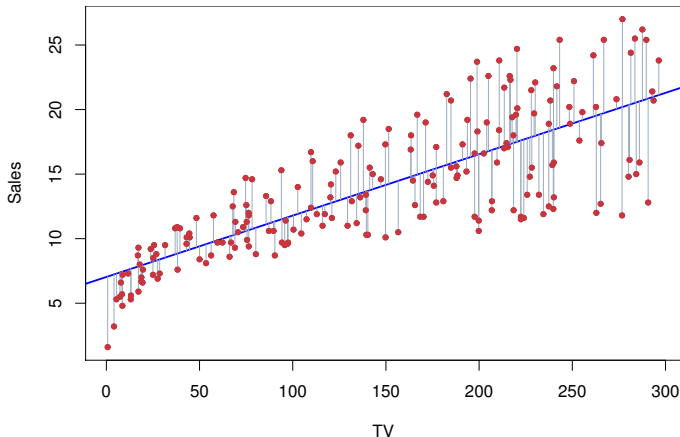
$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$$

$$\hat{\beta}_{0,OLS} = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{1i}$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ are the sample means of x_1 and y .



Ordinary Least Squares (OLS)



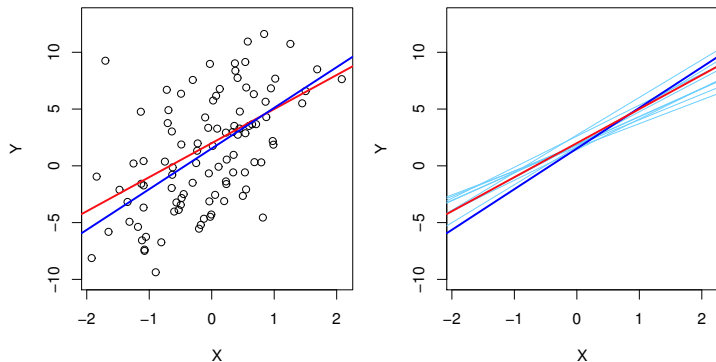
ISLR Fig. 3.1: OLS Linear regression of TV advertising (X_1) on Sales (Y). Note that this is not the *only* way to choose the parameter estimates, but this method provides nice, useful properties.



Note very importantly:

- We think of some “true” / “best” regression line corresponding to the true values β_0 and β_1
 - ▶ This can be the true relationship between X and Y or simply the best linear approximation – in other words, we think of $\hat{\beta}_0$ and $\hat{\beta}_1$ as being estimates of some true values β_0 and β_1 , though these need not describe the true underlying relationship
- However, we *never* observe this: we see only a finite dataset from which we calculate estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
 - ▶ If we got a new dataset, we would get a different estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ and hence a different estimate of the regression line





ISLR Fig. 3.3: **Left:** True regression line (red) and estimate of that line based on the data (blue). **Right:** additional possible estimated regression lines based on different observed datasets (light blue).



- Least squares (OLS) is not the only way we can estimate $\hat{\beta}_0$ and $\hat{\beta}_1$, but it does have several nice properties
- One very nice property: unbiasedness
 - ▶ $\hat{\beta}_0$ is an *unbiased* estimate of β_0 ; we do not systematically over/under-estimate the parameter:

$$\mathbb{E}(\hat{\beta}_{0,OLS}) = \beta_0$$

- ▶ The same holds for β_1

$$\mathbb{E}(\hat{\beta}_{1,OLS}) = \beta_1$$



Inference Procedures: Standard Error

OLS also allows for straightforward calculations of standard errors:

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

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$$

Note: $\sigma^2 = \text{var}(\epsilon)$ which, in practice, we won't know. Instead, we estimate it with the residual standard error (RSE):

$$RSE = \hat{\sigma} = \sqrt{RSS/(n-2)}$$

and write $\hat{SE}(\hat{\beta}_0)^2$ and $\hat{SE}(\hat{\beta}_1)^2$ to indicate the estimate of σ



Inference Procedures: Confidence Intervals

- The standard errors tell us something about how much variability there is in the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- It is common to want to characterize this variability in the form of a *confidence interval* (CI):

$$95\% \text{ CI for } \beta_0 = \hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$$

$$95\% \text{ CI for } \beta_1 = \hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

- **Note:** Book uses a factor of 2, but it should really be ...



Inference Procedures: Hypothesis Testing

- Finally, we'll often want to perform hypothesis tests on β_0 and β_1 ; most commonly

H_0 : No *linear* relationship between X_1 and $Y \iff \beta_1 = 0$

H_1 : Some *linear* relationship between X_1 and $Y \iff \beta_1 \neq 0$

- Under H_0 ,

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

where \sim means “is distributed as” and t_{n-2} denotes the t -distribution with $n - 2$ degrees of freedom

- We can use this result to provide a p-value; small values mean we reject H_0 and conclude that there is some linear relationship between X_1 and Y



- **Big Question:** At the end of the day, we still need to ask the question “How well does the model actually fit our data?”
- A few different ways to assess model fit:

1. Residual Standard Error (RSE):

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

- Large values mean y_i and \hat{y}_i far apart for at least some data
 \implies bad model fit
- Downside:



- **Big Question:** At the end of the day, we still need to ask the question “How well does the model actually fit our data?”
- A few different ways to assess model fit:

1. Residual Standard Error (RSE):

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

- Large values mean y_i and \hat{y}_i far apart for at least some data
 \implies bad model fit
- Downside: Difficult to determine a “good” RSE value because it depends on the scale of Y



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

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

- R^2 = “proportion of variability in Y explained by X_1 in this model”

$\implies 0 \leq R^2 \leq 1$; larger values better

- In this simple linear regression setting, $R^2 = r^2$ where $r = \text{corr}(X_1, Y)$
- Standardizing by TSS means that the measure is invariant to the scale of Y



MULTIPLE LINEAR REGRESSION

Multiple Linear Regression

- Now suppose that instead of only 1 predictor, we have p predictors X_1, \dots, X_p
- The “standard” multiple linear regression (MLR) model is

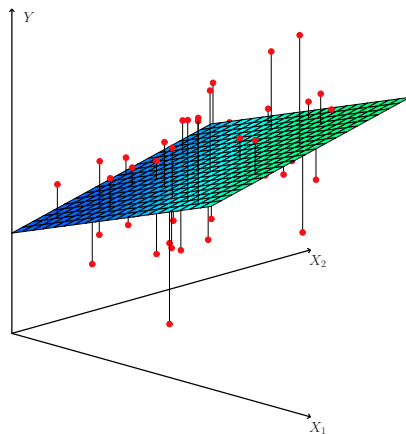
$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

- β_j = average change in Y with 1 unit change in X_j **with all other predictors held fixed**
- Coefficients estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are estimated in same way, by minimizing residual sum of squares:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi}))^2$$



Multivariate OLS



ISLR Fig. 3.4: Visualization of multiple linear regression with two predictors, X_1 and X_2 with the coefficients fit via ordinary least squares (OLS).



Inference Procedures: Hypothesis Testing

- With MLR, we're interested in many of the same kinds of questions as with SLR
- Most simply:
 - H_0 : There is no linear relationship between Y and **any** of the predictors
 - H_1 : There is a linear relationship between Y and **at least one** of the predictors
- Often referred to as the **overall F -test**, since we're testing whether **any** of the predictors are significant



Inference Procedures: Hypothesis Testing

- Equivalent hypotheses:

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

$$H_1: \beta_j \neq 0 \text{ for some } j \in \{1, \dots, p\}$$

- Under H_0 ,

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

where again, \sim means “is distributed as” and $F_{p-1, n-p}$ denotes the F -distribution with $p - 1$ and $n - p$ degrees of freedom (often called the numerator and denominator degrees of freedom, respectively)



Inference Procedures: Hypothesis Testing

- More often we're interested in testing whether a particular *subset* of the predictors are significant
 - ▶ Let X_1^*, \dots, X_m^* denote the subset of predictors we're interested in and $\beta_1^*, \dots, \beta_m^*$ the corresponding coefficients

Hypothesis test becomes

$$H_0: \beta_1^* = \beta_2^* = \dots = \beta_m^* = 0$$

$$H_1: \beta_j^* \neq 0 \text{ for some } j \in \{1, \dots, m\}$$

and under H_0 ,

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

where RSS^* is the residual sum of squares from the linear model that **does not** include X_1^*, \dots, X_m^*



- Recall that when we had only a single predictor (SLR), we could do a t -test for significance of X_1
 - ▶ In fact, even when you perform MLR, statistical software will report a t -test statistic and p-value for each term in the model
- So if we conduct the previous hypothesis test and we're only interested in testing whether one predictor X_1^* is significant, is this the same as just doing the t -test?



- Recall that when we had only a single predictor (SLR), we could do a t -test for significance of X_1
 - ▶ In fact, even when you perform MLR, statistical software will report a t -test statistic and p-value for each term in the model
- So if we conduct the previous hypothesis test and we're only interested in testing whether one predictor X_1^* is significant, is this the same as just doing the t -test?
 - ▶ Yes!



- Recall that when we had only a single predictor (SLR), we could do a t -test for significance of X_1
 - ▶ In fact, even when you perform MLR, statistical software will report a t -test statistic and p-value for each term in the model
- So if we conduct the previous hypothesis test and we're only interested in testing whether one predictor X_1^* is significant, is this the same as just doing the t -test?
 - ▶ Yes!
- So then why can't we just look at each individual t -test, keep the ones that are significant, and throw the rest away?



R Output

```
> str(freeny)
'data.frame':   39 obs. of  5 variables:
 $ y              : Time-Series  from 1962 to 1972: 8.79 8.79 8.81 8.81 8.91 ...
 $ lag.quarterly.revenue: num  8.8 8.79 8.79 8.81 8.81 ...
 $ price.index       : num  4.71 4.7 4.69 4.69 4.64 ...
 $ income.level      : num  5.82 5.83 5.83 5.84 5.85 ...
 $ market.potential  : num  13 13 13 13 13 ...
> lm1 <- lm(y~.,data=freeny)
> summary(lm1)

Call:
lm(formula = y ~ ., data = freeny)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0259426 -0.0101033  0.0003824  0.0103236  0.0267124

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -10.4726     6.0217  -1.739   0.0911 .
lag.quarterly.revenue  0.1239     0.1424   0.870   0.3904
price.index     -0.7542     0.1607  -4.693 4.28e-05 ***
income.level     0.7675     0.1339   5.730 1.93e-06 ***
market.potential  1.3306     0.5093   2.613  0.0133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01473 on 34 degrees of freedom
Multiple R-squared:  0.9981, Adjusted R-squared:  0.9978
F-statistic: 4354 on 4 and 34 DF, p-value: < 2.2e-16

>
```



Example of R output from a linear model.

Two big reasons:

1. Results (coefficient estimates, test statistics, p-values) are *all with respect to all other terms in the model*
 - ▶ If some (at least one) term(s) are removed from the model, the rest changes
2. If each t -test has 5% false positive rate, then about 1 in 20 coefficients will be significant by random chance
 - ▶ If we have 100 predictors, on average, 5 would be significant by chance
 - ▶ F -test adjusts for the number of predictors; doesn't have this problem



Big question then: How do we find the 'best' model (i.e. the model with all significant predictors and no 'extra' redundant information)

- Still very much an ongoing area of research in high-dimensional problems (big data)
- Once we have a given model, many standard statistical models can be used to assess how well that model fits (more on this later in the course)
 - ▶ Akaike Information Criterion (AIC)
 - ▶ Bayesian Information Criterion (BIC)
 - ▶ Mallows's C_p
 - ▶ Adjusted R^2



How do we find good 'candidate' models?

- Given p (potential) predictors, there are 2^p possible models; where do we begin?
 - ▶ **Forward Selection:** Begin with the null model ($Y = \hat{\beta}_0$), add the variable that gives lowest RSS ($Y = \hat{\beta}_0 + \hat{\beta}_1^* X_1^*$), then add the next variable with lowest RSS ($Y = \hat{\beta}_0 + \hat{\beta}_1^* X_1^* + \hat{\beta}_2^* X_2^*$), etc.
 - ▶ **Backward Selection:** Begin with full model ($Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$) and eliminate variables one at a time that have the largest p-value



Inference Procedures: Confidence Intervals

Can create confidence intervals for individual β_i 's, or for $f(X)$, or prediction intervals:

- **95% CI's for β_i :** 95% of these kinds of intervals should cover the true β_i (assuming the truth is really a linear model)
- **95% CI's for $f(X)$:** 95% of these kinds of intervals should cover the true value of $f(X)$ (assuming the truth is really a linear model)
- **95% Prediction Intervals:** 95% of these kinds of intervals should cover the true value of $Y = f(X) + \epsilon$ (assuming the truth is really a linear model)



LINEAR MODEL EXTENSIONS

Thus far we have assumed that the p terms in our model correspond to p continuous predictors X_1, \dots, X_p ; many possible extensions of this setup that still yield linear models:

1. Predictors can be categorical
2. Terms can be functions of individual predictors, e.g.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

3. Terms can be functions of multiple predictors, e.g.

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



Categorical Predictors

1. How do we set up and interpret things if we have a categorical predictor X_i ? Two options:

I. **Dummy Coding:** Treat one level of X_i as the baseline/reference and estimate the change from this baseline for the other levels:

e.g. Let's say we have one continuous predictor X_1 and one categorical predictor X_2 with three levels a_1, a_2 , and a_3 and we treat a_1 as the baseline level:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 1_{x_{2,i}=a_2} + \beta_3 1_{x_{2,i}=a_3} + \epsilon_i$$

$$= \begin{cases} \beta_0 + \beta_1 x_{1,i} + \epsilon_i & \text{if } x_{2,i} = a_1 \\ \beta_0 + \beta_1 x_{1,i} + \beta_2 + \epsilon_i & \text{if } x_{2,i} = a_2 \\ \beta_0 + \beta_1 x_{1,i} + \beta_3 + \epsilon_i & \text{if } x_{2,i} = a_3 \end{cases}$$



Note that with dummy coding, information about the baseline level is incorporated into the intercept term β_0

II. **Effect Coding:** Intercept β_0 is interpreted as the overall mean across all levels of X_2 ; coefficients represent change from this overall mean

e.g. Under the same setup:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 1_{x_{2,i} \neq a_2} + \beta_3 1_{x_{2,i} \neq a_1} + \epsilon_i$$

$$= \begin{cases} \beta_0 + \beta_1 x_{1,i} + \beta_2 + \epsilon_i & \text{if } x_{2,i} = a_1 \\ \beta_0 + \beta_1 x_{1,i} + \beta_3 + \epsilon_i & \text{if } x_{2,i} = a_2 \\ \beta_0 + \beta_1 x_{1,i} + \beta_2 + \beta_3 + \epsilon_i & \text{if } x_{2,i} = a_3 \end{cases}$$



Categorical Predictors

- **Dummy Coding:** β_0 corresponds to the baseline effect of one predictor level
- **Effect Coding:** β_0 corresponds to the overall mean across all predictor levels
- Most software will do dummy coding by default; be careful with this, sometimes it is more intuitive to use one level or another as the baseline
- See ISLR Section 3.3.1 for a more thorough overview on dummy coding; see [here](#) for a good comparison of the methods



2. Terms can be functions of individual predictors

- Usually this takes the form of higher order polynomial terms for one or more of the predictors, e.g.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

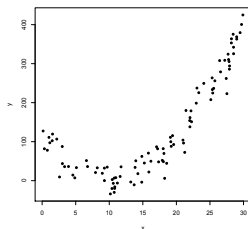
$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2 + \beta_3 X_3^5$$

- Remember, each of these is still linear **in the parameters** so these are still linear models
- In many situations, relationships are clearly not (first-order) linear, but linear models with higher-order polynomial terms may provide a very good fit

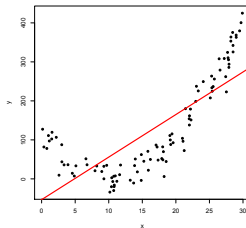


Higher Order Term Example

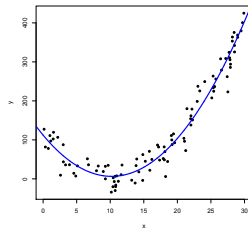
Example of (first order) linear model vs (second order linear) quadratic model



Raw Data



Model Fit: $Y = \beta_0 + \beta_1 X$



Model Fit: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$



3. Terms can be functions of multiple predictors

- Usually this takes the form of products of predictors, e.g.

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

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2 + \beta_3 X_3^5 + \beta_4 X_2 X_3^5$$

- Terms in a linear model that involve multiple predictors are called **interaction terms**. Usually these take the form $X_1^* \cdots X_m^*$ and if the term involves m predictors, we refer to it as an m -way interaction, e.g.

$$X_1 X_2 = \text{2-way interaction} \quad X_1 X_2 X_3 X_4 = \text{4-way interaction}$$

Terms that involve only 1 predictor are called **main effects**



What is implied when we include an interaction term?

- When an interaction term is included and is significant, we say that an **interaction exists** between the predictors involved in that term
- If, for example, the (coefficient corresponding to the) $X_1 X_2$ interaction term is significant, it means the effect/impact of one predictor (say X_1) is different depending on the value of the other (X_2)
- **Note:** Whenever an interaction term is significant, we almost always want to include the individual main effects in the model as well, even if they are not significant



- When an interaction term is included, we still have a linear model but we no longer have an additive model; Additive Models are of the form:

$$Y = f(X) = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) + \epsilon$$

- Models that include only main effects and/or higher order polynomial terms are both additive and linear
- **Note:** In practice, it is rare to include many-way interactions (more than 3-way or 4-way); There are some special circumstances, but generally we only think about 2-way and 3-way interactions



WHAT CAN GO WRONG?

What can go wrong?

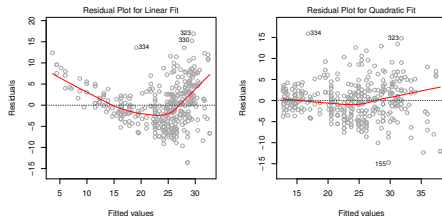
There are a multitude of issues commonly encountered when trying to fit a linear model (see ISLR Section 3.3.3 for more thorough review):

- The true underlying relationship is not linear
- Correlation of error terms
- Non-constant variance of error terms (Heteroscedasticity)
- Outliers and/or high leverage points
- Collinearity/Multicollinearity



What if the true underlying relationship just isn't linear?

- Sometimes including higher order polynomial terms will help; sometimes not. Transforming the response in such a way that makes the transformed relationship more linear can also sometimes help.
- Plots of residuals vs fitted values can help assess non-linearity

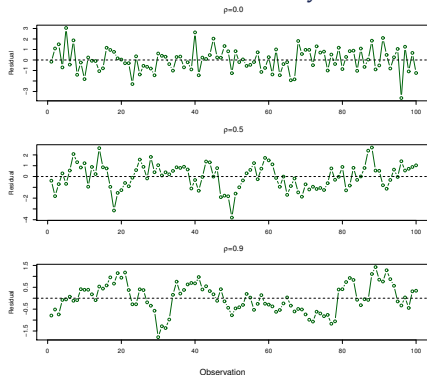


ISLR Fig. 3.9: Example Fitted vs Residual plots. Ideally we want no discernable patterns here.



Correlated Error Terms

- Linear models assume that the error terms (the ϵ_i) are independent, or at least uncorrelated
- Sometimes this is not the case, usually with time series data

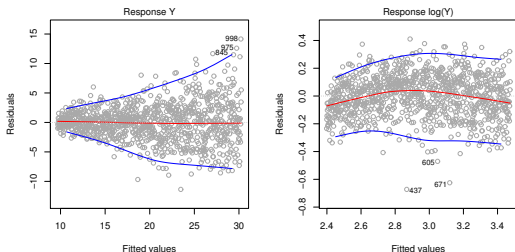


ISLR Fig. 3.10: Residual plots from time series data with varying degrees of correlation ρ .



Heteroscedasticity

- Linear models assume that the error terms (the ϵ_i) have constant variance: $\text{var}(\epsilon_i) = \sigma^2$; note that σ^2 does not depend on i
- Often this is not the case; it is common to see larger error variances for larger predictor values. Transforming the response to $\log(Y)$ or \sqrt{Y} usually helps:



ISLR Fig. 3.11: (Left) Heteroscedastic residual plots using Y as response. (Right) Residual plot using $\log(Y)$ as the response.



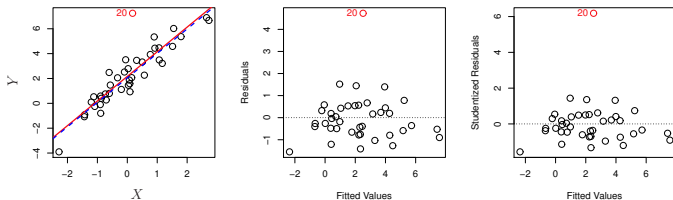
Outliers and High Leverage Points

Outliers and high leverage points are often confused, but are not the same:

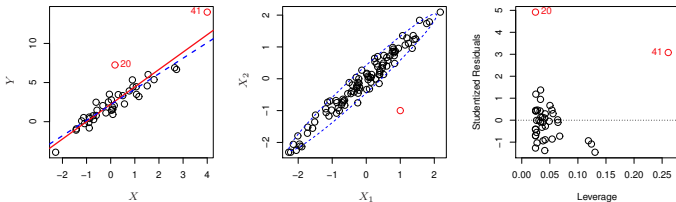
- **Outliers:** Points for which the true value y_i lies far from the predicted value \hat{y}_i
 - ▶ Response y_i is unusual given the predictor value(s) x_i
- **High Leverage Points:** Points for which the value of the predictor(s) x_i lies far from the rest of the data
 - ▶ Often have a substantial effect on model fit
 - ▶ Removing a high leverage point results in a bigger change in \hat{f} than removing an outlier



Outliers and High Leverage Points



ISLR Fig. 3.12: Linear model fit and residual plots. Point 20 is an outlier.

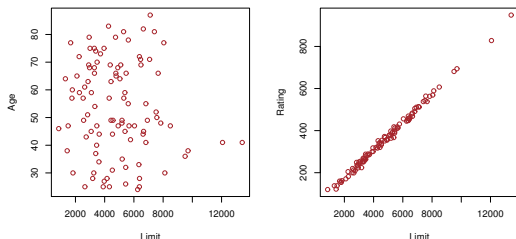


ISLR Fig. 3.13: (Left) Linear model fit; point 20 is an outlier and point 41 is a high leverage point. (Center) Red point indicates an outlier in the case of two predictors X_1 and X_2 . (Right) Outlier and high leverage point in residual plot.



Collinearity and Multicollinearity

- Very common issue; almost always present when many predictors are involved
- Collinearity: very high correlation between predictors to the point where one often explains the other. With only two variables (predictors), it's not *exactly* the same as correlation, but similar (specific meaning in the context of linear regression).



ISLR Fig. 3.14: (Left) Limit and Age are not collinear. (Right) Limit and Rating are collinear.



Effects of collinearity:

- Significant additional uncertainty in estimating the coefficients; significant increase in standard error of the estimates
 - ▶ If we refit the model on new data, we may see very different coefficient estimates
- Suppose X_i and X_j are highly correlated:
 - ▶ When fitting simple linear regression model with either X_i or X_j , both terms look very important
 - ▶ When fitting a multiple linear regression model with *both* terms present, often both look less important (in extreme cases, not even significant)



- Easy enough to spot if we're only dealing with two predictors
 - ▶ Can check before fitting the model by simply looking at covariance matrix
- With more than two predictors, we have to worry about **multicollinearity**
 - ▶ Collinearity that exists between three or more predictors even though pairwise correlations are small
 - ▶ Most common way to check for this is to compute the *variance inflation factor* (VIF) – see book for formula and details



Advantages:

- Simple and explicit model; easy to interpret
- Parameters (β 's) are computationally fast and easy to fit
- Many inference procedures (hypothesis tests and confidence intervals) readily available if you're willing to make certain assumptions

Disadvantages:

- If true underlying relationship between predictors and response is not at least approximately linear, can be some difficulties in interpretation/inference
- Even though a good linear model approximation often exists, it can be *very* difficult to find the 'best' one or even a relatively good one, especially in large dimensions (many predictors)



Bias vs Variance

- Linear models, at least in low-dimensional settings ($p < n$) are generally considered high bias, low variance (not very flexible)
- With linear models, the model “complexity” is determined by the number of terms in the model: more terms \implies more flexible/complex
- Adding more terms to a linear model will improve the error on the training set, but adding useless terms will increase the test error (some very minor exceptions – will discuss in the last lecture of the course)



1. Linear regression can be performed without the strong assumptions generally given in textbooks

- Very few assumptions are needed to fit actually a linear model; stronger assumptions (true underlying linear model, independent Normal errors) are needed for the kinds of inference and interpretations described in the previous slides

2. In practice, there is rarely one “good” linear model

- There will almost always be many different linear models (with different covariates included) with similarly high accuracy – be very careful about reading too much into the particular model you estimated with the data at hand

