Introduction to Statistical Data Science

Thomas Honnor

t.honnor@ucl.ac.uk

Department of Statistical Science, UCL

Linear Regression

Outline

You have studied some of this in the Supervised Learning module. We here present an alternative view that emphasizes interpretation and statistical properties.

Outline

- Basic definitions
- Gaussian vs model-free points of view
- Model checks
- Hypothesis testing and confidence intervals
- Other practical issues

Learning a Relationship

- Our measurements are not independent.
- Often we want to characterize the distribution of an **outcome** Y given observable **covariates X**:

$$Y \sim F(\mathbf{x}, \theta)$$

Outcome of interest: random variable

Some distribution: for instance, Gaussian

"Things we see": covariates, or inputs

The knobs: parameters needed, maybe by estimation. For instance, mean and variance.

Many Names

- These covariates are sometime given many names:
  - Predictors
  - Inputs
  - Regressors
  - Independent variables (bad name!)
- The outcome is sometimes called:
  - Response
  - Output
  - Dependent variable (bad name!)

Learning a Relationship

- But how are parameters related to inputs? We need to specify how they interact to generate *Y*.

$$Y \sim F(\mathbf{x}, \theta)$$

Outcome of interest: random variable

Some distribution: for instance, Gaussian

“Things we see”: covariates, or “inputs”

The knobs: parameters needed, maybe by estimation. For instance, mean and variance.

Example for This Section

- Advertising data (ISLR book).
- Goal: understanding how to improve sales of a particular product.
- Data: sales of that product in 200 markets
  - For each market, budgets spent on TV, radio and newspaper advertisement in thousands of dollars.
  - Outcome: sales, in thousands of units.
  - What is the relationship?

Problem Formulation

- In our problem, Y is sales volume. X, is the advertisement expenditure vector:
  - X_1: TV
  - X_2: Radio
  - X_3: Newspaper
- Task: estimate how Y is related to X.
  - We can apply it to future campaigns, assuming **external validity**: that the relationship in the future remains the same. This can be a strong assumption!

In Matrix Notation

$$\begin{bmatrix} Y^{(1)} & X_1^{(1)} & X_2^{(1)} & X_3^{(1)} \\ Y^{(2)} & X_1^{(2)} & X_2^{(2)} & X_3^{(2)} \\ \dots & \dots & \dots & \dots \\ Y^{(200)} & X_1^{(200)} & X_2^{(200)} & X_3^{(200)} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ \dots \\ Y^{(200)} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & X_3^{(1)} \\ X_1^{(2)} & X_2^{(2)} & X_3^{(2)} \\ \dots & \dots & \dots \\ X_1^{(200)} & X_2^{(200)} & X_3^{(200)} \end{bmatrix}$$

Regression

- Signal + noise:

$$Y^{(i)} = f_{\theta_1}(\mathbf{x}^{(i)}) + \epsilon^{(i)}$$

Signal: the **regression function.**
This is not random (**x** is known)

Error, or “noise”.
This is random

$$\epsilon^{(i)} \sim F(\theta_2)$$

Distribution of error

In what follows, I will typically drop the superscript *(i)* to avoid complicating notation.

Linear Regression with Gaussian Noise

$$Y = \beta_0 + \beta^\top \mathbf{x} + \epsilon$$

- That is,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

and

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

- We have then four free parameters to fit, β_0, β_1, β_2, β_3 and $\sigma^2{}_\varepsilon$.

Why Linear?

- Because it is simple to understand.
- Computationally efficient.
- If you have many variables and not much data, might be as good as it gets (more on that later).
- Don't kid yourself, in most cases reality is not exactly linear, but:
  - **George E. P. Box's dictum, "All models are wrong but some are useful."**

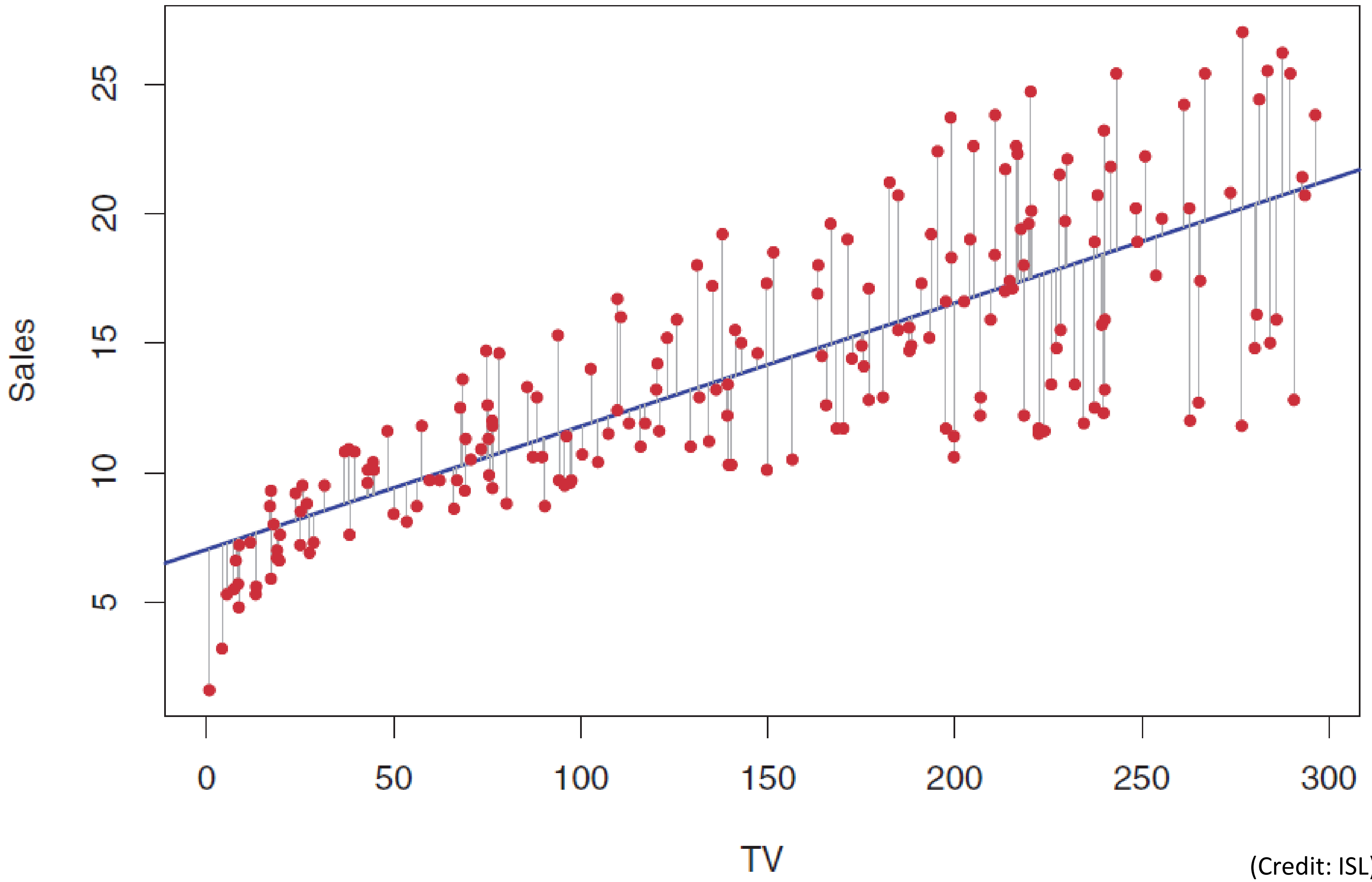
Simple Demo

- One dimensional regression

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

  where *Y* is sales, X_1 is TV budget and the error term here is not the same as in the previous slide (we use the same symbol as an abuse of notation).
- R demo.

The Fitted Model



(Credit: ISL)

Parameter Fitting: What Happened

- What is the model for the data? Recall that

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

- We assume each of our data points $Y^{(1)}$, … $Y^{(200)}$ are independent given **X**.

- They are **not** identically distributed. What is their distribution?

The Distribution of $Y^{(i)}$

- Remember: $\boldsymbol{x}^{(i)}$ here is fixed.
- Each $Y^{(i)}$ is a constant plus some Gaussian random variable $\varepsilon^{(i)}$.
- Without proof, we will claim that $Y^{(i)}$ is Gaussian itself. What are its mean and variance?
- We will use the notation $V = v \mid V' = v'$ to denote the **conditional distribution** of V given V', even if V' is not a random variable.
  – Sometimes $V = v \mid v'$, when V' is obvious from context

Result

- For each data point,

$$Y^{(i)} \mid X_1^{(i)} = x_1^{(i)} \sim N(\beta_0 + \beta_1 x_1^{(i)}, \sigma_\epsilon^2)$$

- Why? Bear with me for a couple of slides.

Mean

- If *Z* is a random variable, what is E[*aZ* + *b*] for two **constants** *a* and *b*?

$$E[aZ+b] = \int (az+b)p(z)dz = a\int zp(z)dz + b\int p(z)dz = aE[Z]+b$$

- So if $Y = \beta_0 + \beta_1 x_1 + \epsilon$,

$$\begin{aligned} E[Y^{(i)} \mid X_1^{(i)} = x_1^{(i)}] &= \beta_0 + \beta_1 x_1^{(i)} + E[\epsilon^{(i)} \mid X_1^{(i)} = x_1^{(i)}] \\ &= \beta_0 + \beta_1 x_1^{(i)} \end{aligned}$$

Variance

- Variance is defined as

$$Var(Z) = E[(Z - E[Z])^2]$$

- That is, nothing but a quantification of how much *Z* differs (in expectation) from its mean by the squared Euclidean distance.
- The use of the name "variance" to describe the scale parameter of a Gaussian wasn't a coincidence.
- We can show that $Var(aZ + b) = a^2 Var(Z)$
- So

$$Var(Y^{(i)} \mid x_1^{(i)}) = 1^2 Var(\epsilon^{(i)}) = \sigma_\epsilon^2$$

Now What?

- If I give you $(\beta_0, \beta_1, \sigma_\epsilon^2)$, you can tell me the probability (density) of each data point.
- We can "play with" the values of these parameters to **maximise the probability of the data occurring**
  - This is essentially the idea we sketched in the previous chapter.
- How to formalize it?

The Likelihood Function

- Probability of the data as a function of parameters:

$$L(\beta_0, \beta_1, \sigma_\epsilon^2) = \prod_{i=1}^{200} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left\{ -\frac{1}{2} \frac{(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)})^2}{\sigma_\epsilon^2} \right\}$$

- It is easier to work on the log-scale, where we also drop constants:

$$\log L(\beta_0, \beta_1, \sigma_\epsilon^2) = -0.5 \sum_{i=1}^{200} \left(\log(\sigma_\epsilon^2) + \frac{(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)})^2}{\sigma_\epsilon^2} \right)$$

The Algorithm

- Now it is a matter of computing the maximum of this likelihood function.
- This will be given the too-obvious name of **maximum likelihood estimator (MLE)**.
- Finding a MLE can be computationally hard in general (more about that in future chapters!), but here it can be done analytically.
  - That is, take derivatives, set them to zero, solve equations.

Result

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{200}(x_1^{(i)} - \bar{x}_1)(y^{(i)} - \bar{y})}{\sum_{i=1}^{200}(x_1^{(i)} - \bar{x}_1)^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 \\
\hat{\sigma}_\epsilon^2 &= \frac{1}{200}\sum_{i=1}^{200}(y^{(i)} - \hat{\beta}_0 \; - \hat{\beta}_1 x_1^{(i)})^2
\end{aligned}
$$

where

$$
\bar{y} = \frac{1}{200}\sum_{i=1}^{200} y^{(i)} \qquad \bar{x}_1 = \frac{1}{200}\sum_{i=1}^{200} x_1^{(i)}
$$

Prediction

- Now for every point x_1, we can provide a **prediction** for Y at that point.

- As in Chapter 1, we can think of the conditional expectation as an appropriate prediction.

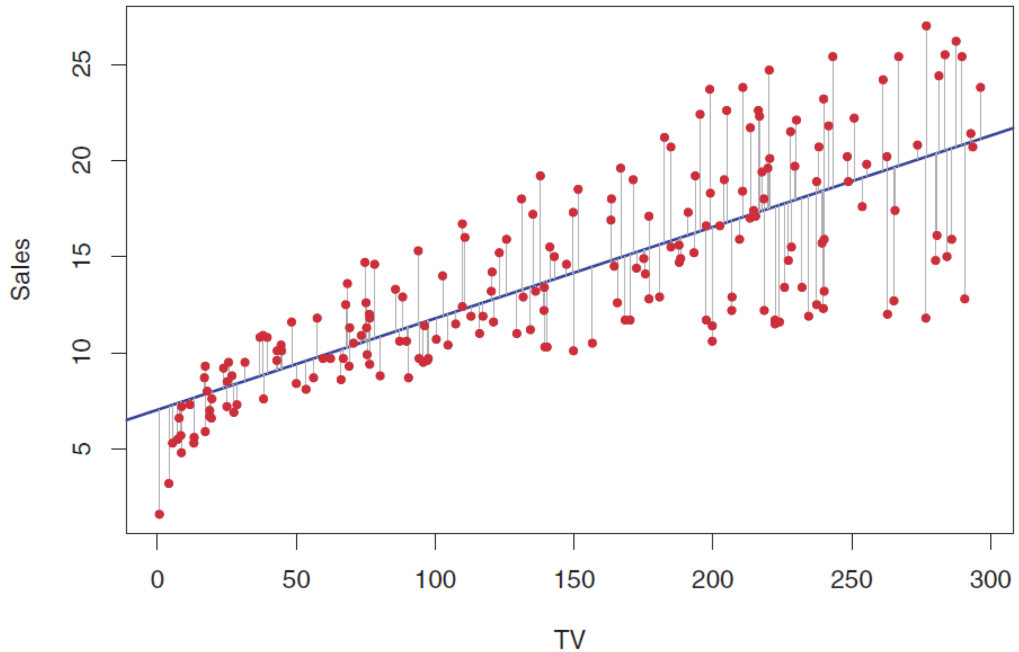
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Prediction

- Using terminology from the Machine Learning literature, we call the data we used to fit the model the **training data**.
- It is common to reserve some data to evaluate how well we can perform **out-of-sample**, that is, with future unseen data. This data we reserved is called **test** (or testing) **data**.
  - There are more sophisticated ways of partitioning the data between training/test. See the Supervised Learning class (also, some in Chapter 5).

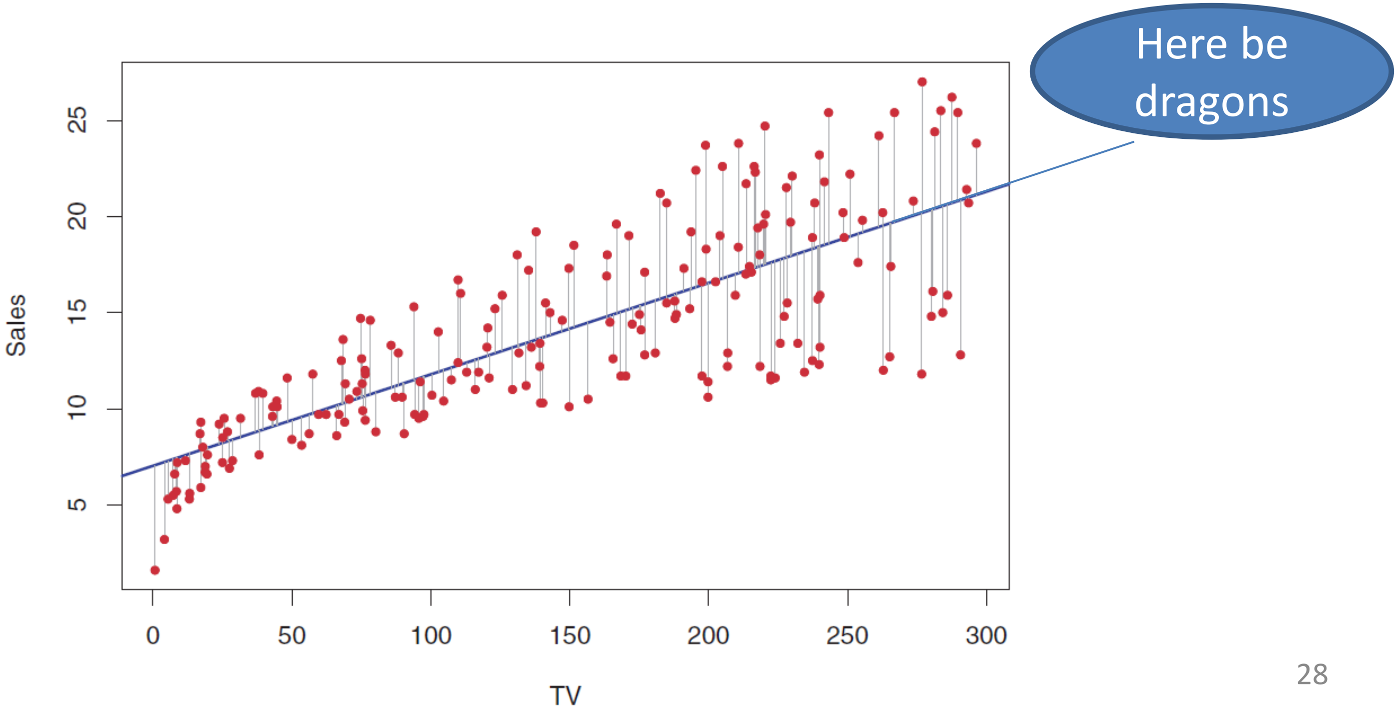
Smoothing

- We can think of **smoothing** as a way of "denoising" the **training data** you had. It provides estimates of the expectation **within-sample.**



Extrapolation

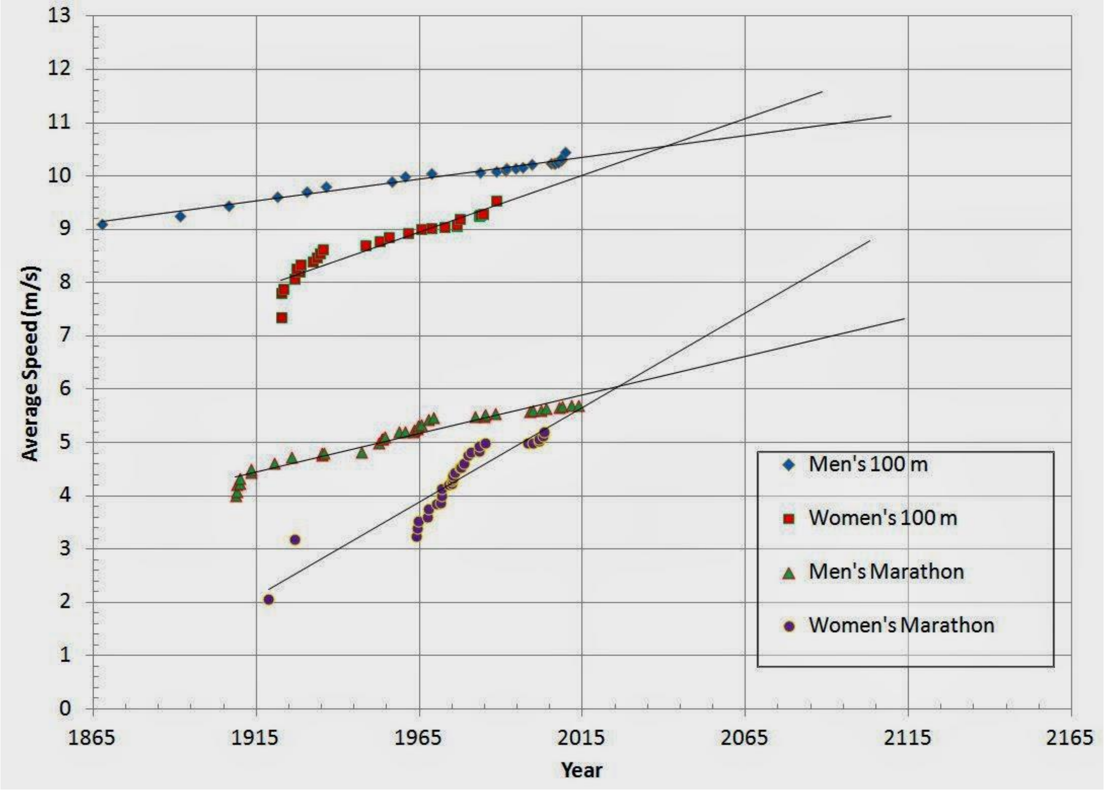
- Predictions “outside” the training data.
- Not always easy to define.
- Beware of unwarranted extrapolations!



Extrapolation

- Particularly tempting with linear models.

"This is not me talking, it's the data."



http://www.nytimes.com/1992/01/07/science/2-experts-say-women-who-run-may-overtake-men.html

Extrapolation

If she loves you more each and every day,
by linear regression she hated you before you met.

http://www.smbc-comics.com/comic/2011-08-05

Diagnostics

- Is this a good model? Bad? In which ways?
- Which kind of visual checks can we have as the number of inputs grows to 2, 3, …, very many?
- Ways in which we can get things wrong:
  - Non-linearity!
  - Noise is not "homogenous" (heteroscedasticity)
  - Non-Gaussianity?
- In what follows, n will be used to denote sample size and p will denote number of inputs.

REGRESSION WITHOUT GAUSSIANITY ASSUMPTIONS

Non-Gaussianity

We will first discuss why it is not necessary to assume Gaussianity, and why and when we do/don't.

Residuals

- What you miss by your linear reconstruction.

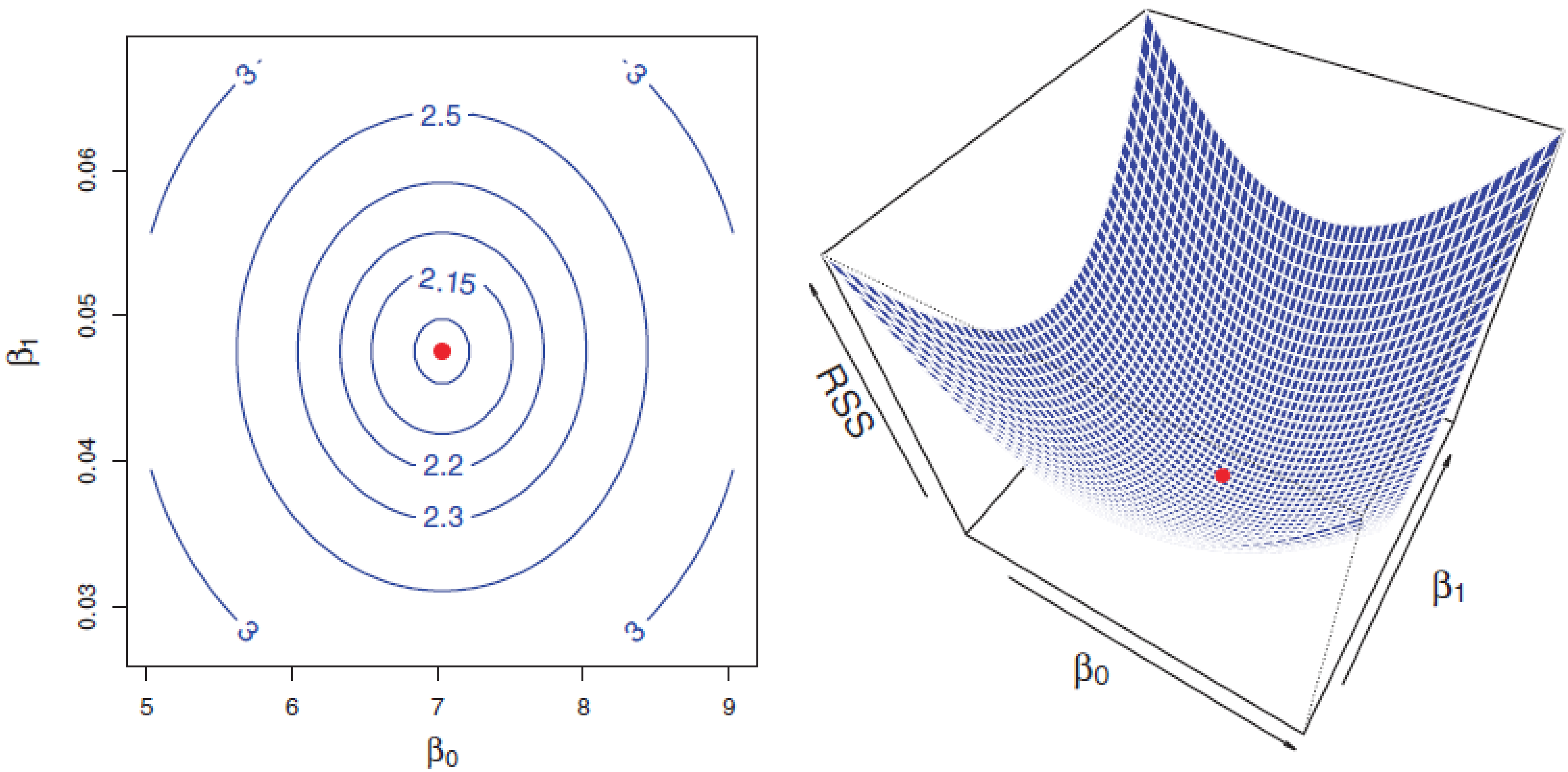
$$e^{(i)} \equiv y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)}$$

- Summary: **residual sum of squares (RSS)**

$$RSS \equiv \sum_{i=1}^{n} {e^{(i)}}^2$$

- **What is the relation between that and the log-likelihood function?**

Least-Squares Interpretation



$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y^{(i)} - \beta_0 - \beta_1 x_i^{(1)})^2$$

Least-Squares Estimator

This is identical to the Linear Gaussian MLE for the regression coefficients β.

What Does it Mean?

- Say you have a sample of independent, identically distributed (**i.i.d**) random variables

$$Y^{(i)} \sim F(\theta)$$

  for i = 1, 2, …, n. Say you want to estimate the mean of this distribution by maximum likelihood.
- What would you do for Gaussians?

MLE for iid Gaussians

- If we maximise this

$$\log L(\mu, \sigma^2) = -0.5 \sum_{i=1}^{n} \log(\sigma_\epsilon^2) + \exp\{(y^{(i)} - \mu)^2/\sigma^2\}$$

we get the **sample mean**

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

- Notice: this is just a special case of Gaussian linear regression with an empty set **X**.

What If We don't Want to Assume Gaussianity?

- Isn't this intuitive?

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

- Of course it is, but how to justify it? Enter the **empirical cdf** again.

$$\hat{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} I(x^{(i)} \leq x)$$

Empirical cdf vs Population cdf

- A central result of **nonparametric statistics** is that the empirical cdf converges (in a probabilistic sense I won't define) to the population cdf as *n* grows

$$\hat{F}_n(x) \to F(x)$$

- If you must know, this is called the Glivenko-Cantelli theorem.
- R demo.

Nonparametrics?

- I won't say much about this now, except that these are statistical models we can't describe with a finite number of parameters
  - More on that in Introduction to Supervised Learning, and later chapters.
- Suffices to say that unlike the Gaussian model that uses two parameters, the empirical cdf estimate uses *n* "parameters". It is not a fixed number.

How Do We Use This?

- Expectation, now with the "**empirical pdf**"!

$$E[Y] = \int y\hat{p}(y)dy = \sum_{i=1}^{n} y^{(i)}\frac{1}{n} = \bar{y}$$

- Implication? The sample mean is justified as a **consistent** estimator of means
  – It "converges" to the truth, as *n* increases
- This happens without assuming Gaussianity
  – even if it is exactly the same formula as in the Gaussian case

Implications to Regression

- Gaussianity assumption is not necessary to estimate the regression function, or even the error variance.
- However, we cannot say anymore that we can estimate the conditional distribution

$$Y \sim F(\mathbf{x}, \theta)$$

- So, if you need something other than mean/variance, you will need further assumptions such as Gaussianity.
- And if the conditional distribution is "far" from Gaussian, don't fool yourself that least-squares will be reliable.

Concluding This Discussion

- We can do statistical inference without likelihood functions.
  - Bayesian inference requires likelihood function, see *STATG004*.
- There are important advantages in likelihood modelling
  - Full uncertainty modelling.
- However, more assumptions.
  - Keep in mind though that generality is not the same as reliability.

RESIDUAL ASSESSMENT AND MODEL CHECKS

Residual Assessment and Model Checks

Now we assess what residuals can tell us about accuracy, non-linearity and heteroscedasticity.

R^2 Statistic

- The RSS is not that straightforward to interpret because of its scale.
- R^2 is a proportion statistic. More exactly, the proportion of variance of Y explained by **X**. It always take values between 0 and 1.

$$R^2 \equiv \frac{TSS - RSS}{TSS}$$

where

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

$$TSS = \sum_{i=1}^{n} (y^{(i)} - \bar{y})^2$$

(Total sum of squares)

Interpretation

- *TSS* – *RSS* measures "amount of variability in the outcome that is explained".
- If we get 0, the linear model does not provide a good explanation for the data.
- Let's check the R^2 of our advertisement example using R.
  – R also includes something called "adjusted R^2". The difference matters little for large sample sizes.

High $R^2 \neq$ Good Predictions

- High R^2 is good news, but may be not enough.
- Although in theory regression doesn't make assumptions about the distribution of covariates **X**, its interpretation will require assumptions. This includes interpreting R^2.
  - Recall the talk about extrapolation
- R demo with synthetic data.

High $R^2 \neq$ Good Predictions

$$R^2 = \frac{a^2 Var(X)}{a^2 Var(X) + Var(\epsilon)}$$

- This means $R^2 \rightarrow 0$ as Var(X) $\rightarrow 0$ and $R^2 \rightarrow 1$ as Var(X) $\rightarrow \infty$!
- Even with much non-linearity we can get high R^2!
- In particular, bad predictions with high R^2 will follow if Var(ε) is high, but Var(X) is much higher.
  - "bad" in the sense of absolute error, not necessarily in the sense of relative error with respect to not having seen **x**.

High $R^2 \neq$ Good Predictions

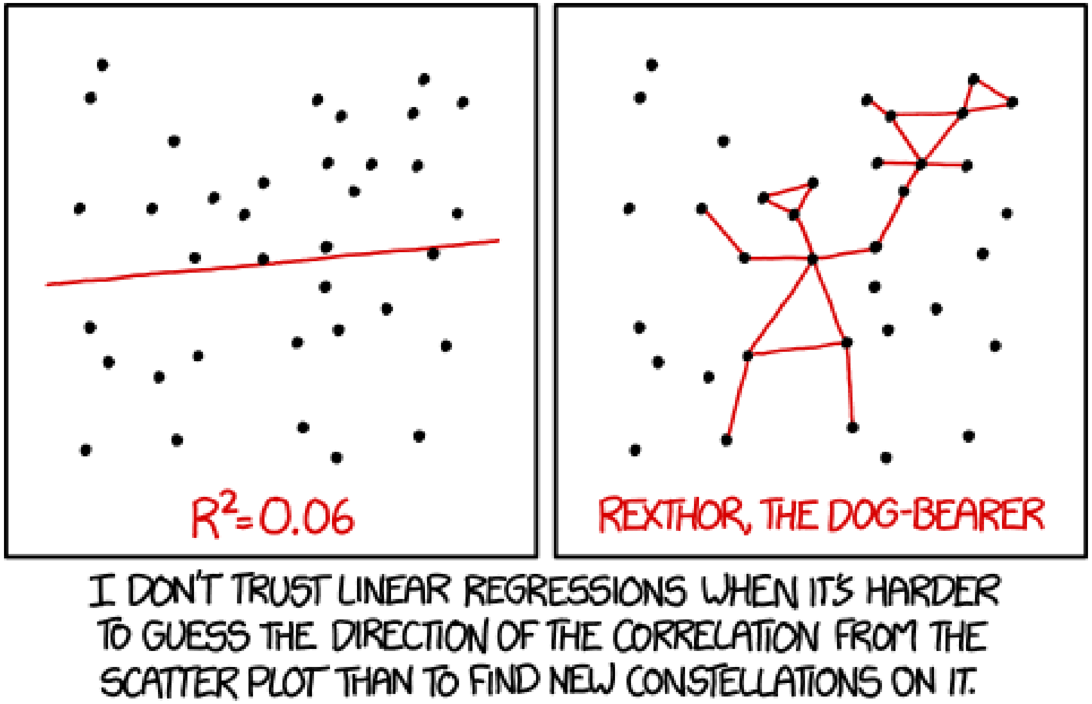
- Despite that, it is good practice to report R^2, as a "necessary but not sufficient" diagnostic of how good your fit is.
- (Assuming fixed model) However, there is only so much we can achieve with a given **x**:

$$\begin{aligned} E(Y - \hat{Y}^2) &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\epsilon) \end{aligned}$$

We can shrink this with better modelling

To shrink this, we may need to measure further variables

In Any Case…



http://xkcd.com/1725/

Residual Plots

- What should we expect to see in a good regression model?

$$e^{(i)} \equiv y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)}$$

- R demo: in what follows we will exemplify diagnostics by comparing the advertising model to the outcome of a well-behaved synthetic model.

Residual Plots

- R's *lm* plot 1: residuals vs fitted

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \qquad e^{(i)} \equiv y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)}$$

- What you should expect to see is lack of correlation between the two. In particular
  - The location (empirical average) and spread (empirical variance) of the residual axis stays similar across the value of the fitted outcomes.

Residual Plots

- With this plot, it might be possible to detect **outliers**, points “far from the curve” that may or may not indicate model failure.
  - Notice that the scale will depend on Y.
  - It could be the natural result of non-Gaussian error, for instance.
  - It could be the result of measurement error that *maybe* should be removed.
  - At this stage, we won’t be formal about outliers. One thing to keep in mind at this time, however, any outlier removal should be documented and justified.

Residual Plots

- R's *lm* plot 2: Normal Q-Q
  - As we have seen before, the assumption of normality is not necessary.
  - However, if there are highly skewed residuals, you might want to ask whether the mean of the outcome is a good estimand to target.
  - Violations of normality have other implications to model checking, to be discussed later.

Residual Plots

- R's *lm* plot 3: Scale-Location
  - Similar to plot 1, but transformed: square roots of absolute value of standardised residuals.
    - "standardised" = divided by empirical standard deviation
  - Rationale: horizontally, should show "no pattern" (flat red line, homogenous spread around it)
    - For Gaussian errors: approximately most points should be less than 2. But main point it to visualize homogeneity.
    - Square root is just to minimize the visual impact of more extreme points.

Residual Plots

- R's *lm* plot 4: Residuals vs. leverage
- Think of a concept that complements regression outliers: while outliers refer to point "off the *y* axis", it measures instead how points are "off the bulk of *x* values".
- This is straightforward to visualize in one dimension. For higher dimensions, we reduce it to a single number, **leverage**, which summarizes it.

Residual Plots

- R demo: let's compare two synthetic datasets that differ only by one data point.

- In one-dimension, the leverage statistic for data point x_i is

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

Residual Plots

- The value of the leverage statistic is always is between $1 / n$ and 1.
- If we have p inputs, the average leverage across inputs is $(p + 1) / n$.
- Values deviating “much” from this average can be flagged.

Residual Plots

- In the residual vs. leverage plot, a point can be an **outlier** (large standardised residual) and/or a **high leverage** point.

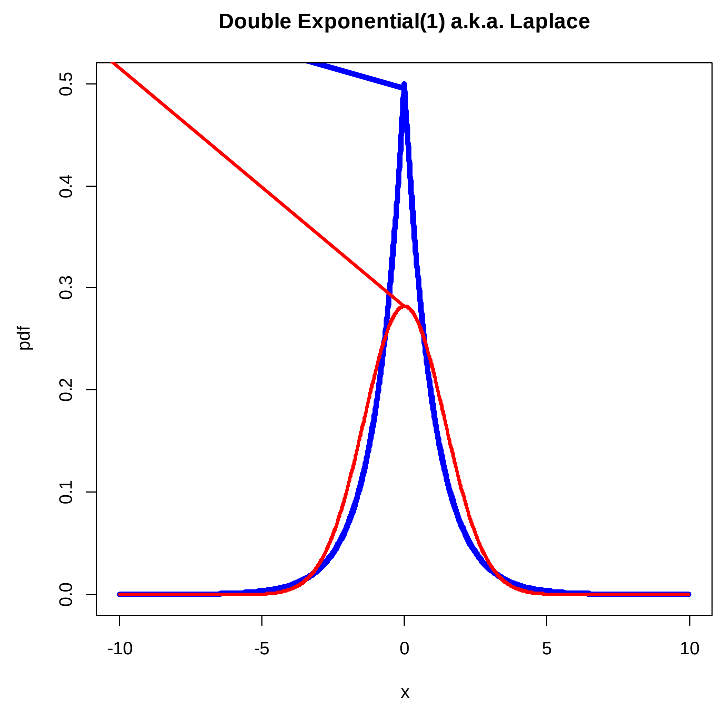
- In R, the vertical axis is standardised, but the horizontal axis is relative. So the point of highest leverage may be inconsequential anyway.

R Demos

- Now let’s walk though these plots again for an idealized examples contaminated with outliers, and one with high leverage points.

R Demos

- Now let’s walk though these plots again for an idealized example with errors which follow a **double-exponential** (a.k.a. **Laplace**) distribution



In the figure, blue is the density of a Laplace(1), red is N(0, 2).

Both of these distributions have variance 2.

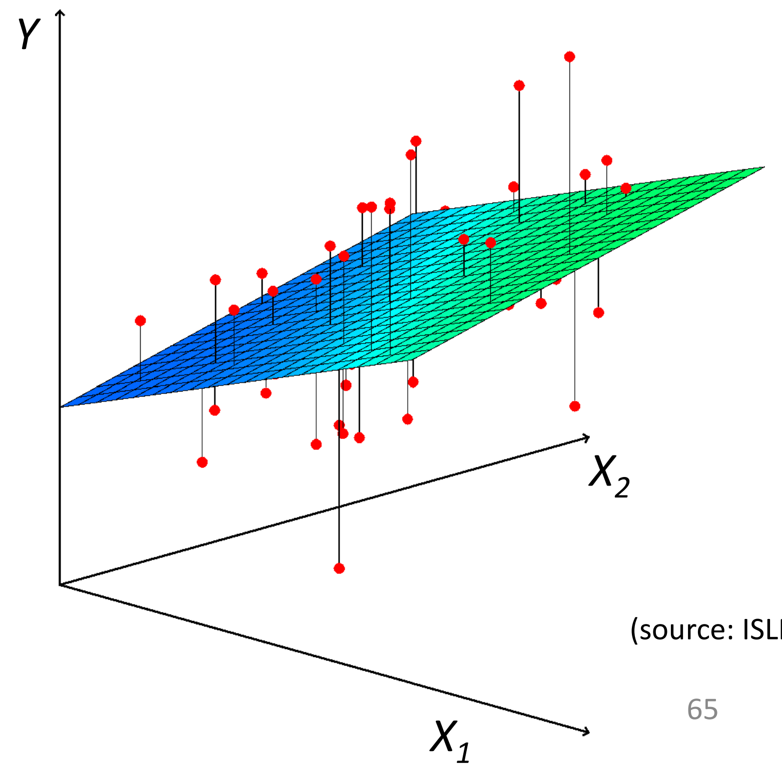
R Demos

- Now let's walk though the plots again for our advertisement data.

Finally: Multiple Regression

- Let's just fit this model, where X_1, X_2 and X_3 are budgets for TV, radio and newspaper, respectively (R demo).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$



(source: ISLR)

HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

Two Basic Null Hypothesis

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

- Rejecting it would mean β_i at least one coefficient is non-zero.

- That is, is there any association of any kind between input Y and X_i given the other inputs?
  - Notice the subtle "given the other inputs". More on that in the future.

The All Zero H_0

- For the former hypothesis, consider the following statistic

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

which refers to our old friends

$$RSS = \sum_{i=1}^{n} (y^{(i)} - \hat{y}(i))^2 \qquad TSS = \sum_{i=1}^{n} (y^{(i)} - \bar{y})^2$$

How Can *F* Falsify H_0?

- If you feel like doing some algebra, you will be able to show that

$$E\left[\frac{RSS}{n-p-1}\right] = \sigma^2$$

where σ^2 is the variance of the error term.

- Under the null the following is also true:

$$E\left[\frac{TSS-RSS}{p}\right] = \sigma^2$$

- So, what would you suggest?

A Test of H_0

- The *F* statistic should be "close" to 1 under the null.
- We have a machinery to decide what closeness is:
  - Find the distribution of *F*
  - Assess the probability (with respect to the data distribution) that *F* is greater than 1
    - Technical note: $E\left[(TSS - RSS)/p\right] > \sigma^2$ if H_0 is false.
  - Reject H_0 if this probability is smaller than your agreed test level (sigh… "0.05" for the sake of illustration)

Implications

- If your p-value is low, H_0 is rubbish at explaining the data: reject it.
  - If you must know, the *F* statistic follows (approximately, in the non-Gaussian case) the unimaginatively named *F* distribution, which I won't explain.
- This is a test that is commonly reported, **but don't fool yourself that this is evidence of a good model**.
  - Your data is arguably very bad if this very strong H_0 is not rejected.

R Demo

- Recall that our linear model of sales volumes looks preposterous. But guess its p-value.

Testing Subsets

- There are analogous F statistics for the null $\beta_i = 0$ only (that is, the other coefficients are unconstrained). As a matter of fact, we can easily test whether any subset of coefficients is zero.

- Many software packages report the one-coefficient test automatically.

Implications

- If your p-value is high, there is evidence predictor X_i does not explain the variability of the outcome *given the other predictors*.

- This is not the same as input X_i not being important.

- R demo.

Implications

- If you do find evidence that predictors are important (e.g., tests give low p-values), again this does not mean the model is "good".

- However, testing provides an useful indication of which variables are redundant or not quite useful, *given the sample size you have and the model assumptions.*
  - They *might* prove useful if you later collect larger sample sizes.

Beware of the Star-Chasing Complex

```
Call:
lm(formula = adv$Sales ~ adv$TV + adv$Radio + adv$Newspaper)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.938889   0.311908   9.422   <2e-16 ***
adv$TV         0.045765   0.001395  32.809   <2e-16 ***
adv$Radio      0.188530   0.008611  21.893   <2e-16 ***
adv$Newspaper -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,     Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

These might be there even if your model has no predictive value

Beware of the Star-Chasing Complex

- In a later chapter we will discuss variable selection and what it means in practice.

Confidence Intervals

- This is very similar to the general idea. Find some pivot around which an interval can be built.

- As a technical aside, let us define an adjusted estimator for the error variance.

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \beta_1 x_1 - \cdots - \beta_p x_p)^2$$

(from now on, we assume β_0 can be represented by $\beta_1 \times 1$, where X_1 is always 1.)

Confidence Intervals

- The following can be shown to be true when errors are Gaussian:

$$T_i \equiv \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{v_{ii}}} \sim \mathcal{T}(n-p)$$

  and v_{ii} is the *i*th entry of the diagonal of $(\mathbf{X}^T\mathbf{X})^{-1}$.
- Notice this requires $n > p$
  – As a matter of fact, least-squares is ill-defined if $n < p$.
- Exercise: write an expression for a confidence interval for β_1 of coverage $1 - \alpha$.
- Note: CLT applies and Gaussianity ends up again not being that important.

Predictive Intervals

- A different matter is **predictive intervals**.

- In Supervised Learning, you may see a lot about prediction. But going one step further, we might want to characterize uncertainty in the prediction. This takes into account uncertainty of the estimates.

What We mean by That

- If we assume a model like the Gaussian, then this implies uncertainty as a conditional distribution

$$Y = \beta_0 + \beta_1 x_1 + \epsilon \Leftrightarrow Y \mid X_1 = x_1 \sim N(\beta_0 + \beta_1 x_1, \sigma^2)$$

- First, let's look at the uncertainty of the **expected value of outcome** given inputs when all we have are parameter estimates.

Prediction

- Say a new data point x_1^* comes, and you want to predict the output as follows

$$\hat{Y}^\star \equiv \hat{\beta}_0 + \hat{\beta}_1 x_1^\star$$

- We know the coefficients themselves are random variables if we consider the training data to be random. What is the long-run variability of my prediction?

Answer

$$Var(\hat{Y}^\star) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_1^\star)$$

- Let's not worry about how to calculate this. The important message is the interpretation: **the randomness here is in the estimated coefficients**, and they come from the randomness in the training data.
- To get the variance of Y^* itself (notice the lack of a hat), we also use the variance of the error.

Predictive Variance

- In practice, we cheat a little bit: the estimated variance $\hat{\sigma}_\epsilon^2$ of the error term is given by the empirical variance of the residuals and treated *as if* it was known.
- Also, recall Var(W_1 + W_2) = Var(W_1) + Var(W_2) for two arbitrary **independent** random variables W_1 and W_2.
- So

$$Var(Y^\star) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_1^\star + \epsilon) \approx Var(\hat{\beta}_0 + \hat{\beta}_1 x_1^\star) + \hat{\sigma}_\epsilon^2$$

(R demo)

OTHER PRACTICAL ISSUES AND DIAGNOSTICS

Interpretation of Regression Models

- We fit the model of sales volume against TV, radio and newspaper expenditures. We get this:

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon$$

- What is its interpretation?

  – Recall first the units: sales volume is measured in thousands of units; each advertising budget is in thousands of dollars.

Interpretation of Regression Models

- The dangerous conclusion:

"If we increase the TV budget by one thousand then, other things being equal, I will sell 400 hundred more units of my product, in expectation."

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon$$

Careful!

- Let me tell you the following extra piece of information: this data came from an **observational study**.
- That is: there was no documented explanation on the causes leading to the level of TV expenses.
- Why this is relevant? Because there might be **common causes (confounding)** of both sales volume and TV expenses. **They may be hidden.**
  - For instance, TV budgets are bigger in markets that are stronger economically, where also people are more likely to buy your product anyway.

Careful!

- Under some strong assumptions, it might be possible to extract causal effects from observational studies. In other situations, you might have **randomized controlled trials**. Then your regression coefficients can be interpreted as causal effects.
  - A big topic in itself that I will leave entirely to *STATG002*
- Without these conditions, some people still refer to regression coefficients as “effects”. This is common, but I find it preposterous.

Interpretation of Regression Models

- A more sober conclusion:

  "Other budgets being equal, an increase of TV budget by one thousand dollars will correspond to an increase of 400 hundred more units of my product, in expectation."

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon$$

- **Notice the major difference**: "increase" here means a increase "as in" the training set, whatever black-box mechanism that was.

Interpretation of Regression Models

- "Other budgets being equal". **Your regression coefficients depend entirely on which other variables are included.**
- Notice the major difference!

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon_{123}$$

$$Y = 12.35 + 0.05x_3 + \epsilon_3$$

I'm emphasizing which variables I'm using as inputs.

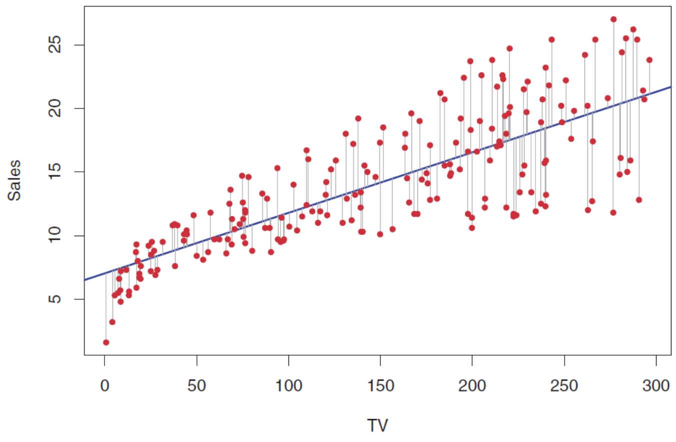
- Be careful to contextualize what you mean by a variable being "important".

The Linear Elephant in the Room

- We know that for our advertising data, linearity is not particularly great.
- There are all sorts of great nonlinear black-box models
  - Introduction to Supervised Learning, and Chapter 5 of our course, will address very many of them.
- However, it sometimes pays off to improve the humble linear model with a change of representation.

Logarithm Transforms

- Heteroscedasticity looks strong in this problem.
- Sometimes it is the result of **multiplicative errors**.



$$Y = x\epsilon$$

- Logarithm transforms can be taken with non-negative data.

Logarithm Transforms

- In our advertising data, let's try taking
  - the logarithm of TV budget
  - sales volume
  - both
- R demo.

Well, That Wasn't Great Was it?

- But it illustrates the principle that we can stick to a linear model that builds a nonlinear mapping (logarithm, in this case).
  - A principle that is taken to the extreme with kernel methods, as discussed in Supervised Learning.
- Other transformations can be done, for instance using a quadratic polynomial.

$$Y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1^2 + \epsilon$$

Interactions

- From the point of view of interpretability, one common use of the linear model is through quadratic or higher order polynomials, e.g.
- This is in part to the idea of interpreting **interactions**. For instance, with the advertising data, is there a "synergy" effect between media?
  - Spending more money on radio could "change the slope" for TV, if the extra exposure makes people pay more attention to TV adverts.

Interactions

- In a linear model, this is translated by constructing inputs derived from the product of more basic inputs.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

- R demo.
- Notice that pairwise interactions already substantially increase the number of inputs.

Notice

- By adding non-linear transformations as inputs to your linear model, this essentially gives you a test of whether the linear model in the original space was reasonable.

- Because if the hypotheses of zero-coefficient for the non-linear terms are rejected, then the original representation was not good enough.

Discrete Inputs and Interpretation

- In Supervised Learning, you may have seen already how to deal with discrete inputs.

  – In Statistics, sometimes we use the generic term **categorical variable** or **factor** to mean discrete variable. Discrete variables can also be **ordinal** if they have a meaningful ordering (think number of stars in a Netflix rating), and can, of course, be **counts**.

Discrete Inputs

- Binary variables (e.g., gender) can be typically represented as 0 or 1, as we have seen before.
- In the context of regression

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

translates to either

$$Y = \beta_0 + \beta_1 + \epsilon \qquad \text{or} \qquad Y = \beta_0 + \epsilon$$

Example

- The *Credit* dataset (from ISLR).
- "Balance" as output, "Gender" as input. We will treat (arbitrarily) level *Female* as 1, *Male* as 0. Let's do a R demo.
- Alternatively, we could code these levels as 1 and -1:

$$Y = \beta_0 + \beta_1 + \epsilon \qquad\qquad Y = \beta_0 - \beta_1 + \epsilon$$

so β_0 can be interpreted as the "baseline credit balance"

Interpretation with More than Two Levels

- We can again create a "dummy" encoding, again with the idea of having one fewer variable than the number of values.
  - So, one dummy for binary variables, two for variables with three levels and so on.
- The reason for the "one fewer" rule is the lack of **identifiability** otherwise.
  - That is, there are infinitely many coefficients giving the same output.

Interpretation with More than Two Levels

- For instance, let's say we have X_1 as an indicator that someone is male (x_1 = 0 if person is not male, 1 otherwise). Let's have X_2 as an indicator that someone is female.
  - Clearly $x_1 + x_2 = 1$, so these two models are identical for any c:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$Y = (\beta_0 - c) + (\beta_1 + c)x_1 + (\beta_2 + c)x_2 + \epsilon$$

Interpretation with More than Two Levels

- This restricted encoding is not symmetric, but without it the linear model would break down.
  - R demo.

- The choice of “base level” is up to the practitioner.

- See also: **analysis of variance** in *STATG002*.

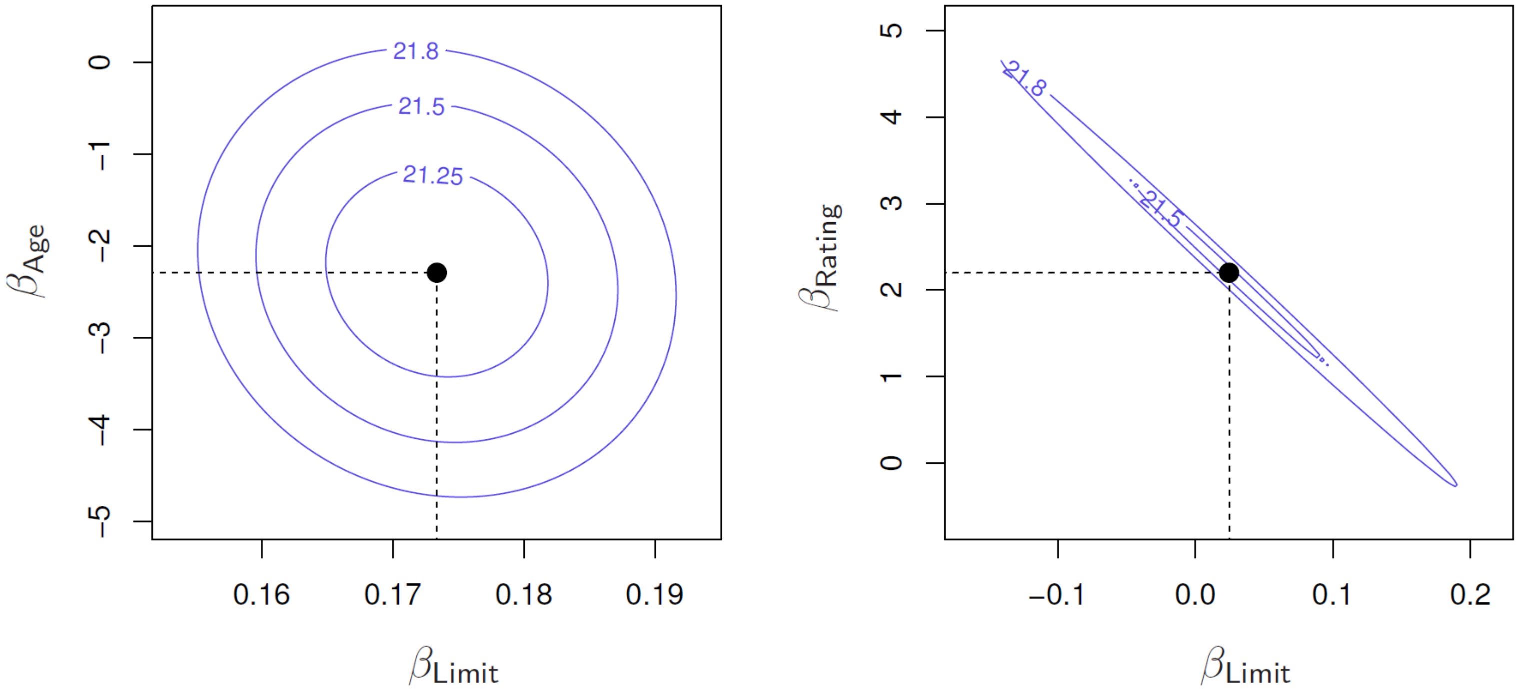
Final Comment: Collinearity

- A “softer” version of the problem of identifiability in linear models: variables which are almost linear combinations of others.

- What happens in the Credit dataset? For instance, the relation between credit and rating? (R demo)

Collinearity

- How does the RSS change with parameter values? Bivariate regression plots.



(ISLR)

Collinearity

- Interpreting parameters of variables which are linearly related is not possible due to unidentifiability.

- Interpreting parameters of variables which are almost linearly related may be unreliable due to wide confidence intervals.

Collinearity

- Try to understand whether it makes sense to have nearly-collinear variables in your model. Remember that each coefficient describes the association between a given input and output holding the other inputs fixed.
  - This is “mutually assured destruction” if an input can basically be derived from the others.

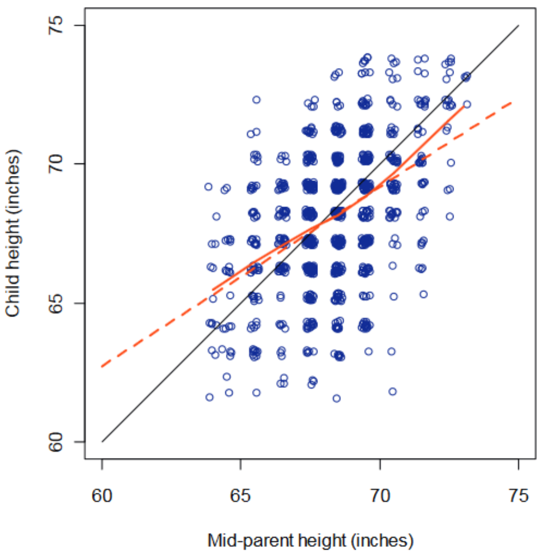
Take-Home Messages

- Linear regression is *the* workhorse of data analysis.
- Prediction is not everything: judicious interpretation of the model and where it fails to fit is also there to convey further messages.
- Confidence intervals help with that, while hypothesis testing provides some basic evidence of what your data can tell about the model components.

Next: model-based regression beyond Gaussianity.

A Historical Note

- The idea of least-squares dates back to at least Gauss and Legendre.
- The name “regression” itself comes from Francis Galton.



Yes, the very same Galton who names our lecture theatre. He was a mentor of Karl Pearson, who founded our department.

S. Stigler (1981). “Gauss and the invention of least-squares”. http://projecteuclid.org/euclid.aos/1176345451

S. Senn (2011). “Francis Galton and the regression to the mean”. http://www.dcscience.net/Senn-2011-Francis-Galton-and-Regression-to-the-Mean.pdf