Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 4 due Friday 10/13

Midterm

Introductions

- One question or thought related to lecture last week (Inference in multiple regression, linear regression through matrices, transformations and assumptions)

The case of the disappearing significance

This section will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here.

```
countries <- read.csv("data/countries.csv")
```

A few useful columns:

- `spi_ospi`: Overall social progress index on 0-100 scale
- `mad_gdppc`: GDP per capita
- `wdi_expedu`: Government expenditure on education as percent of GDP
- `wvs_fight`: Proportion of people who answered affirmatively to "Of course, we all hope that there will not be another war, but if it were to come to that, would you be willing to fight for your country?"
- `wvs_jabribe`: Average response on scale of 1 (never justifiable) to 10 (always justifiable) to the question: "Do you think the following can be justified: someone accepting a bribe in the course of their duties."

1. Fit four linear models to predict a country's overall social progress index: one based on each variable. Interpret each result. Do these make sense?

```
summary(lm(spi_ospi ~ mad_gdppc, countries))$coefficients
```

```
##                Estimate   Std. Error  t value     Pr(>|t|)
## (Intercept) 5.986646e+01 1.297101e+00 46.15403 6.279206e-91
## mad_gdppc   4.656548e-04 4.621341e-05 10.07618 1.392096e-18
```

```
summary(lm(spi_ospi ~ wdi_expedu, countries))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 47.616452  4.1062185 11.596181 2.877746e-21
## wdi_expedu   4.908073  0.8867178  5.535101 1.882912e-07
```

```
summary(lm(spi_ospi ~ wvs_fight, countries))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  92.96888   5.589175 16.633739 1.036614e-26
## wvs_fight   -24.36110   7.859978 -3.099385 2.741961e-03
```

```
summary(lm(spi_ospi ~ wvs_jabribe, countries))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 87.900880   4.121866 21.325505 2.501228e-33
## wvs_jabribe -6.319874   2.096233 -3.014872 3.520518e-03
```

GDP per capita is significantly positively associated (t-stat 10.1, p-value $\leq 10^{-15}$) with a country's overall social progress index with a \$10000 increase in GDP per capita associated with a 4.7 point improvement in the OSPI. Education expenditures are significantly positively associated (t-stat 5.5, p-value $\leq 10^{-6}$) with a country's overall social progress index with a 1% GDP increase in education spending associated with a 4.9 point improvement in the OSPI. Willingness to fight for one's country is significantly negatively associated (t-stat -3.1, p-value .003) with a country's overall social progress index with a 10% increase in citizens willing to fight associated with a 2.4 point decrease in the OSPI. Acceptability of a bribe is significantly associated (t-stat -3.0, p-value 0.004) with a decrease in a country's overall social progress index with a 1 point increase in bribe acceptability corresponding to a 6.3 point decrease in OSPI. These all seem about right, although the "willingness to fight" effect might be a little surprising.

2. Fit a linear model to predict a country's overall social progress index from GDP per capita and education expenditures. Find the correlation between the coefficients for GDP and education expenditures.

```
lm2 <- lm(spi_ospi ~ mad_gdppc + wdi_expedu, countries)
summary(lm2)
```

```
##
## Call:
## lm(formula = spi_ospi ~ mad_gdppc + wdi_expedu, data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.927  -6.647   2.634   7.312  15.232
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.665e+01  3.300e+00  14.137  < 2e-16 ***
## mad_gdppc   4.447e-04  4.958e-05   8.970 7.82e-15 ***
## wdi_expedu  3.407e+00  7.312e-01   4.659 8.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.85 on 112 degrees of freedom
##   (79 observations deleted due to missingness)
## Multiple R-squared:  0.5304, Adjusted R-squared:  0.522
## F-statistic: 63.26 on 2 and 112 DF,  p-value: < 2.2e-16
```

```
# This is a pretty good general purpose command for getting the correlation of predictors
# Divides the variance covariance matrix by the standard deviation of the rows and columns
# to get the pairwise correlations
lm2_vcov <- vcov(lm2)
t(t(lm2_vcov/sqrt(diag(lm2_vcov))) /
    sqrt(diag(lm2_vcov)))[2:nrow(lm2_vcov), 2:nrow(lm2_vcov)]
```

```
##             mad_gdppc wdi_expedu
## mad_gdppc   1.0000000 -0.2188587
## wdi_expedu -0.2188587  1.0000000
```

Every \$10000 increase in GDP per capita is associated with a 4.5 point increase in OSPI after controlling for the association with education expenditures, and the result is very significant (t-stat 9.0, p-value $< 10^{-14}$). A 1% increase in percent GDP spent on education is associated with a 3.4 point increase in OSPI after controlling for the association with GDP per capita, and the result is very significant (t-stat 4.7, p-value $< 10^{-5}$).

3. Fit a linear model to predict a country's overall social progress index from all the variables we have considered so far and print the summary. What happens to the `wdi_expedu` significance? Print

the pairwise correlation between the fit coefficients. What's explaining the change in `wdi_expedu` significance? How can you test this?

```
lm3 <- lm(spi_ospi ~ mad_gdppc + wdi_expedu + wvs_fight + wvs_jabribe, countries)
summary(lm3)
```

```
##
## Call:
## lm(formula = spi_ospi ~ mad_gdppc + wdi_expedu + wvs_fight +
##     wvs_jabribe, data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.857  -2.734   2.131   4.309   8.174
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.924e+01  6.631e+00  11.950  < 2e-16 ***
## mad_gdppc    4.871e-04  6.635e-05   7.342 9.42e-10 ***
## wdi_expedu   1.152e+00  7.059e-01   1.632  0.10835
## wvs_fight   -1.584e+01  5.355e+00  -2.959  0.00452 **
## wvs_jabribe -4.261e+00  1.545e+00  -2.759  0.00782 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.459 on 56 degrees of freedom
##   (133 observations deleted due to missingness)
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7322
## F-statistic: 42.02 on 4 and 56 DF,  p-value: 3.023e-16
```

```
lm3_vcov <- vcov(lm3)
t(t(lm3_vcov/sqrt(diag(lm3_vcov))) /
    sqrt(diag(lm3_vcov)))[2:nrow(lm3_vcov), 2:nrow(lm3_vcov)]
```

```
##              mad_gdppc  wdi_expedu  wvs_fight wvs_jabribe
## mad_gdppc    1.0000000 -0.48876078  0.3952057  0.33689404
## wdi_expedu  -0.4887608  1.00000000 -0.1256962  0.04314546
## wvs_fight    0.3952057 -0.12569615  1.0000000  0.39439134
## wvs_jabribe  0.3368940  0.04314546  0.3943913  1.00000000
```

```
# Checking to see if the dropped observations explain the change
lm4 <- lm(spi_ospi ~ mad_gdppc + wdi_expedu,
          countries[!is.na(countries$wvs_fight) & !is.na(countries$wvs_jabribe),])
summary(lm4)
```

```
##
## Call:
## lm(formula = spi_ospi ~ mad_gdppc + wdi_expedu, data = countries[!is.na(countries$wvs_fight) &
##     !is.na(countries$wvs_jabribe), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.362  -2.132   2.115   5.232   8.301
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 5.875e+01  3.090e+00  19.011  < 2e-16 ***
## mad_gdppc   5.874e-04  6.434e-05   9.129 8.14e-13 ***
## wdi_expedu  1.013e+00  7.530e-01   1.345    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.981 on 58 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.6976, Adjusted R-squared:  0.6872
## F-statistic: 66.91 on 2 and 58 DF,  p-value: 8.623e-16
```

```
lm4_vcov <- vcov(lm4)
t(t(lm4_vcov/sqrt(diag(lm4_vcov))) /
    sqrt(diag(lm4_vcov)))[2:nrow(lm4_vcov), 2:nrow(lm4_vcov)]
```

```
##             mad_gdppc wdi_expedu
## mad_gdppc   1.0000000 -0.5183139
## wdi_expedu -0.5183139  1.0000000
```

Education spending goes from being very significant to being insignificant. The pairwise correlations are not very strong between education spending and the new variables added, but the pairwise correlation has increased considerably between education and GDP per capita. Looking carefully, we can see that the number of included observations has decreased from 115 to 61, so almost half of the data points have been dropped. When we refit the original model with this subset of the data, we see that the eliminated data was driving the significance before.

4. Seychelles is an archipelagic country consisting of 115 islands in the Indian Ocean off the east coast of Africa. It has a GDP per capita and education expenditure percentage listed but no OSPI. Use the GDP per capita and education expenditures model to provide a 95% confidence and prediction interval for OSPI. What is each interval trying to capture?

```
predict(lm2,newdata = countries[countries$cname=="Seychelles",],interval = "confidence")
```

```
##          fit      lwr      upr
## 154 74.84227 72.57096 77.11359
```

```
predict(lm2,newdata = countries[countries$cname=="Seychelles",],interval = "prediction")
```

```
##          fit      lwr      upr
## 154 74.84227 53.22061 96.46394
```

The confidence interval is intended to capture the average of all OSPIs for countries like Seychelles. The prediction interval is intended to capture the actual OSPI for a country like Seychelles.

Contrast test and limiting cases

Recall the setup for a contrast test: $H_0 : \vec{C}^T\vec{\beta} = \gamma_0$ vs. $H_a : \vec{C}^T\vec{\beta} \neq \gamma_0$. Under the null, the following random variable has a $t_{n-(p+1)}$ distribution.

$$T = \frac{\vec{C}^T\hat{\vec{\beta}} - \gamma_0}{\hat{\sigma}\sqrt{C^T(X^TX)^{-1}C}}$$

1. Name two situations in which we would take γ_0 to be 0.

We could take γ_0 to be 0 if we wanted to see whether a single predictor had any effect. We could also take γ_0 to be 0 if we wanted to see whether the difference in two contrast vectors was significant.

2. Write a function that will take a linear model and two named vectors and run a contrast test for a difference between the response variable. Hint: `vcov(fit_lm)` is the same as $\hat{\sigma^2}(X^TX)^{-1}$ and

`coef(fit_lm)` gets the fit coefficeints of the model.

```
contrast.test <- function(fit_lm, vec1, vec2) {
  beta.hat = coef(fit_lm)
  C = vec1 - vec2
  t.stat = C %*% beta.hat/sqrt(t(C) %*% vcov(fit_lm) %*% C)
  p.value = 2*(1-pt(abs(t.stat),df=fit_lm$df.residual))
  return (c("t.stat" = t.stat, "p.value" = p.value, "df" = fit_lm$df.residual))
}
```

3. Perform a formal contrast test based on the GDP per capita plus education expenditures model to determine whether the mean OSPI for (mythical) countries like Seychelles is significantly different from the mean OSPI for (mythical) countries like Madagascar.

```
c("(Intercept)" = 1, unlist(countries[countries$cname == "Seychelles",
                                      c("mad_gdppc", "wdi_expedu")])) -
  c("(Intercept)" = 1, unlist(countries[countries$cname == "Madagascar",
                                        c("mad_gdppc", "wdi_expedu")]))
```

```
## (Intercept)   mad_gdppc  wdi_expedu
##     0.00000 28103.46021     1.59553
```

```
contrast.test(lm2, c("(Intercept)" = 1, unlist(countries[countries$cname == "Seychelles",
                                                         c("mad_gdppc", "wdi_expedu")])),
              c("(Intercept)" = 1, unlist(countries[countries$cname == "Madagascar",
                                                    c("mad_gdppc", "wdi_expedu")])))
```

```
##    t.stat   p.value        df
##  11.14118   0.00000 112.00000
```

To determine whether OSPI in countries like Seychelles is higher than in countries like Madagascar based on GDP per capita and education expenditures, we want to test $H_0 : 28103\beta_1 + 1.6\beta_2 = 0$ vs $H_a : 28103\beta_1 + 1.6\beta_2 \neq 0$. The resulting t-statistic is 11.1 based on 112 degrees of freedom, resulting in a p-value of $< 10^{-5}$, so we reject the null and conclude that countries with predictors like Seychelles are very likely to have a higher average OSPI than countries with predictors like Madagascar.

4. Name and check two limiting cases to ensure the function works as intended.

- When two contrast vectors differ by 1 in a single predictor, we expect to see the same result as when running a t-test on that predictor itself.
- When the model contains only a single categorical predictor, a test between a contrast vector with both 1s vs a contrast vector with only the intercept as 1 should be the same as running a t-test.

```
# Make sure single predictor case collapses to t-test
contrast.test(lm2, c("(Intercept)" = 0, "mad_gdppc" = 1, "wdi_expedu" = 0),
              c("(Intercept)" = 0, "mad_gdppc" = 0, "wdi_expedu" = 0))
```

```
##       t.stat      p.value           df
## 8.970127e+00 7.771561e-15 1.120000e+02
```

```
summary(lm2)$coefficients
```

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 4.664802e+01 3.299638e+00 14.137315 1.161577e-26
## mad_gdppc   4.447532e-04 4.958159e-05  8.970127 7.822117e-15
## wdi_expedu  3.406503e+00 7.312413e-01  4.658521 8.827898e-06
```

```
# Binary predictor t-test
lm5 <- lm(spi_ospi ~ as.factor(bmr_dem), countries)
```

```
summary(lm5)$coefficients
```

```
##                      Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)          57.14557   1.489484 38.366008 6.687378e-83
## as.factor(bmr_dem)1 18.83948   1.971916  9.553899 2.052877e-17
```

```
contrast.test(lm5, c("(Intercept)" = 0, "bmr_dem" = 1),
              c("(Intercept)" = 0, "bmr_dem" = 0))
```

```
##     t.stat    p.value         df
##   9.553899   0.000000 161.000000
```

```
t.test(spi_ospi ~ bmr_dem, countries)
```

```
##
##  Welch Two Sample t-test
##
## data:  spi_ospi by bmr_dem
## t = -9.5534, df = 148.73, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -22.73628 -14.94268
## sample estimates:
## mean in group 0 mean in group 1
##        57.14557        75.98505
```

The t-test from a linear model and the contrast test on a single coefficient always give the same result as expected. A t-test on the data itself and a contrast test for a categorical variable's impact give every so slightly different results. I'm not sure whether this is due to numerical precision issues, degrees of freedom estimation, or something else. (Later realization is that it has to do with the number of items in each group. If the groups are balanced the t-statistics are the same and the only thing different is the degrees of freedom.)

P-values high and low

1. Explain intuitively how increasing the number of predictors in a model could reduce the significance of a particular predictor relative to a model with only that predictor. Find a pair of matrices X_1 and X_2 that illustrate this point. Hint: Recall that the standard error of $\hat{\beta}_i$ is $\hat{\sigma}\sqrt{C^T(X^TX)^{-1}C}$ where C is a one-hot encoded vector with the i^{th} entry as a 1.

We have observed this phenomenon a few times already: this is often caused by predictors with a large amount of collinearity increasing $(X^TX)^{-1}$. For example, for $X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $(X^TX)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ but for $X = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $(X^TX)^{-1} = \begin{bmatrix} 2.22 & -1.78 \\ -1.78 & 2.22 \end{bmatrix}$, so the diagonals (what contribute to the standard error) increase.

2. Explain intuitively how increasing the number of predictors in a model could increase the significance of a particular predictor relative to a model with only that predictor.

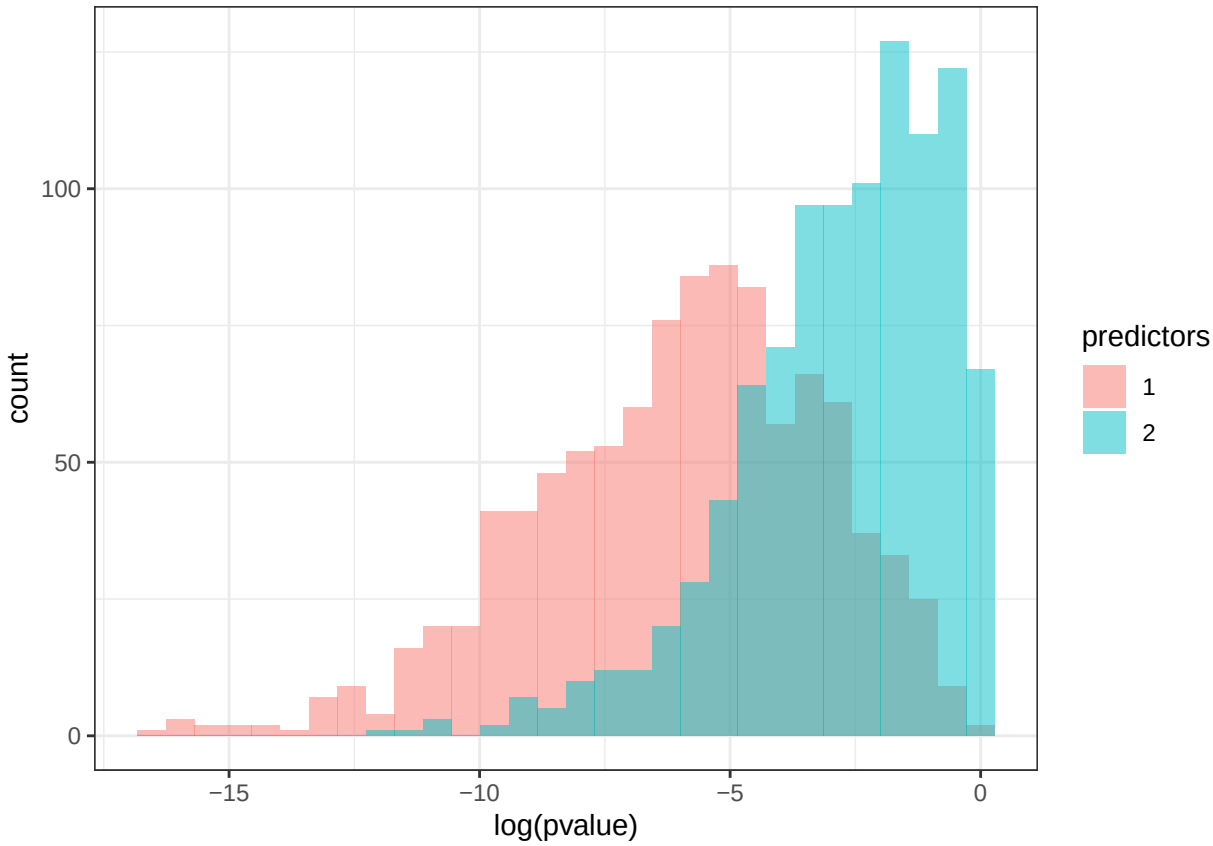
If the additional predictors are useful for prediction and they are orthogonal to the original predictor's observed values, the additional predictors can decrease $\hat{\sigma}^2$, increasing the significance.

3. Interpret the results of the following simulation:

We will generate data from the model $Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Let $\beta_1 = 1$, $\beta_2 = 2$, $\sigma^2 = 2^2$.

To show how the significance can decrease, we will first have $\vec{X}_1$ and $\vec{X}_2$ be correlated and test models with and without X_2.

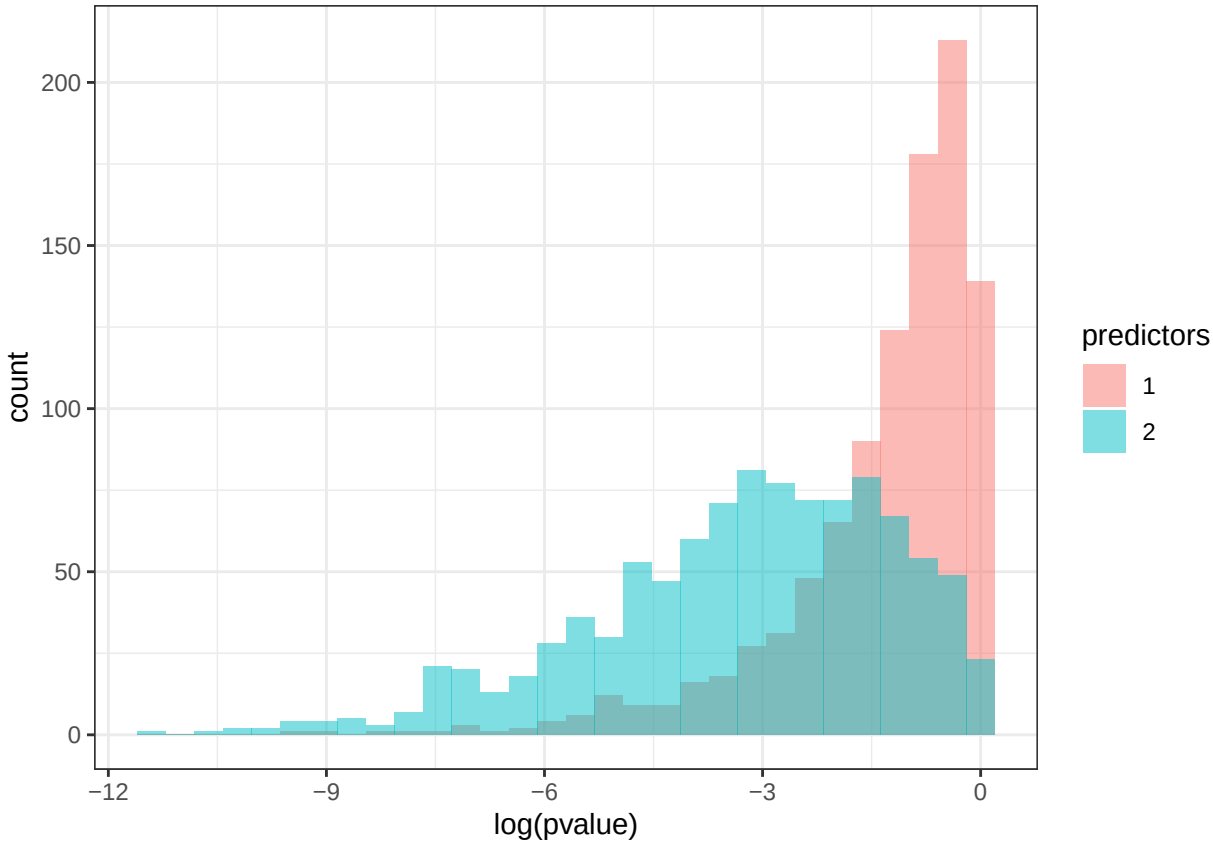
```
library(MASS)
library(ggplot2)

nsims = 1000
n = 20
beta_1 = 1
beta_2 = 2
sigma = 2

# Correlation matrix for Xs
Sigma = cbind(c(1, 0.5), c(0.5, 1))

# Model with predictors for only X_1
pvals_single = vector(length = nsims)
for (i in 1:nsims) {
  x = mvrnorm(n = n, rep(0, 2), Sigma)
  y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
  pvals_single[i] = summary(lm(y ~ x[,1]))$coefficients[2, 4]
}

# Model with predictors for X_1 and X_2
pvals_double = vector(length = nsims)
for (i in 1:nsims) {
  x = mvrnorm(n = n, rep(0, 2), Sigma)
  y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
  pvals_double[i] = summary(lm(y ~ x))$coefficients[2, 4]
}

out_data = data.frame(predictors = as.factor(c(rep(1, nsims), rep(2, nsims))), pvalue = c(pvals_single,

ggplot(out_data, aes(x = log(pvalue), fill = predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw()
```



To show how significance can increase, we will have $\vec{X_1}$ and $\vec{X_2}$ be uncorrelated but have X_2 have a very large variance, and we will test models with and without X_2.

```
nsims = 1000
n = 20
beta_1 = 1
beta_2 = 2
sigma = 2

# Notice the very large variance of X_2
Sigma = cbind(c(1, 0), c(0, 10))

# Model with X_1 as the only predictor
pvals_single = vector(length = nsims)
for (i in 1:nsims) {
  x = mvrnorm(n = n, rep(0, 2), Sigma)
  y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
  pvals_single[i] = summary(lm(y ~ x[,1]))$coefficients[2, 4]
}

# Model with both X_1 and X_2 as predictors
pvals_double = vector(length = nsims)
for (i in 1:nsims) {
  x = mvrnorm(n = n, rep(0, 2), Sigma)
  y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
  pvals_double[i] = summary(lm(y ~ x))$coefficients[2, 4]
}
```

```
out_data = data.frame(predictors = as.factor(c(rep(1, nsims), rep(2, nsims))), pvalue = c(pvals_single,

ggplot(out_data, aes(x = log(pvalue), fill = predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins = 30) +
  theme_bw()
```

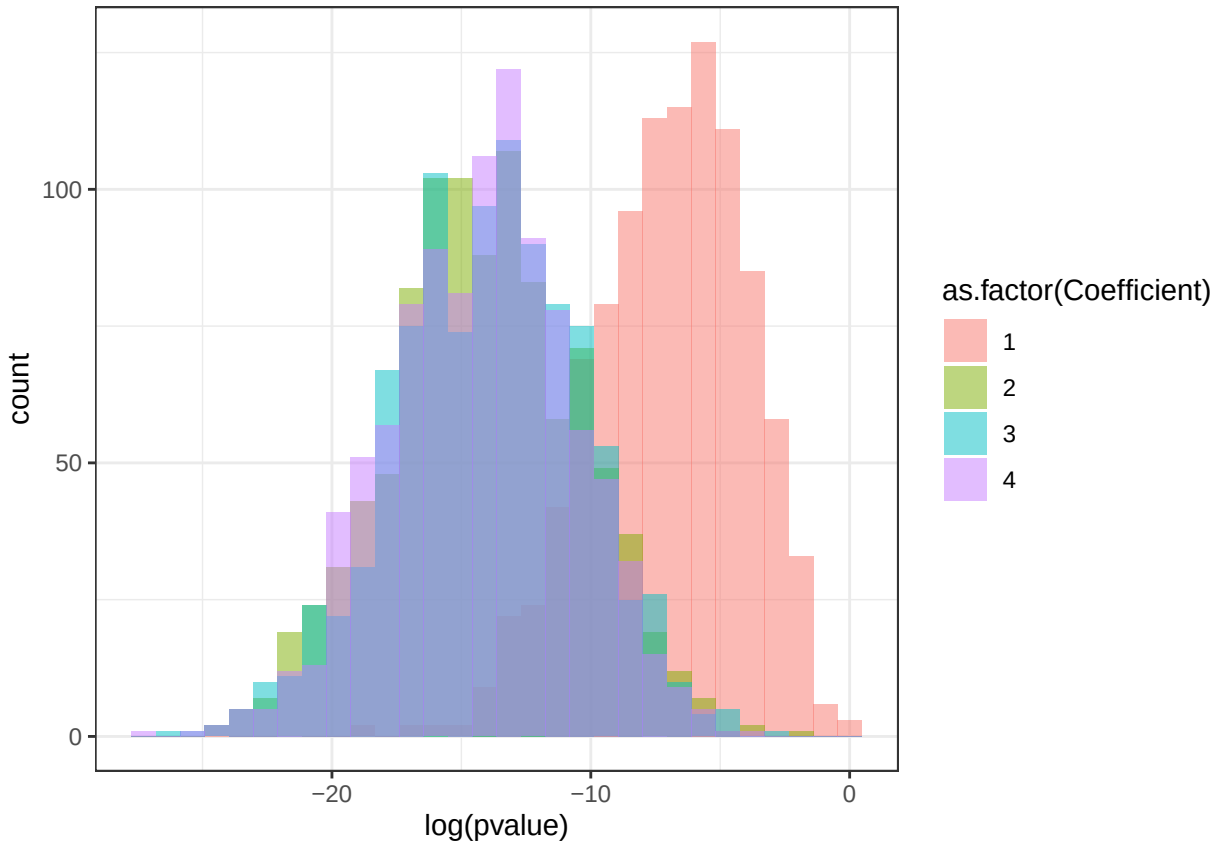


Winner takes all

1. Many times in the real world we have a situation where some particular predictor underlies many other predictors. The following simulation shows a situation in which there is an underlying variable `x_main` and three related predictors `x_2`, `x_3`, and `x_4`. First look at the histograms from when the coefficients and variances are the same for all the Xs. Compare a multiple regression model with a single-variable regression model for each coefficient and comment on what's happening. Then, by changing the values of the beta coefficients and the variances of `x_main`, `x_2`, `x_3`, and `x_4`, explore what factors drive significance in a multiple regression model.

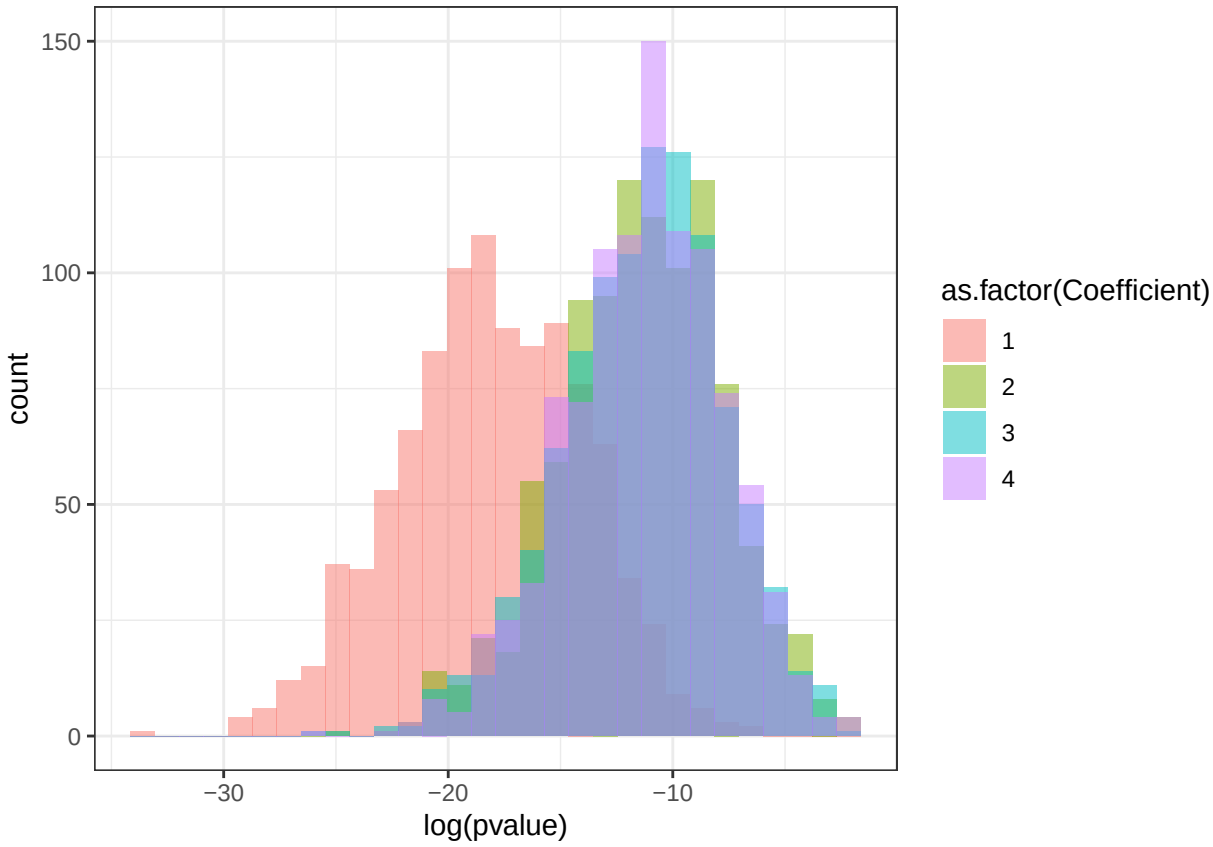
```
library(reshape2)

nsims = 1000
n = 20
beta_1 = 2
beta_2 = 2
beta_3 = 2
beta_4 = 2
sigma = 2
```

```
# Model with X_1 as the only predictor
pvals = matrix(nrow = nsims, ncol = 4)
for (i in 1:nsims) {
  x_main = rnorm(n, 0, 2)
  x_2 = rnorm(n, x_main, 2)
  x_3 = rnorm(n, x_main, 2)
  x_4 = rnorm(n, x_main, 2)
  y = beta_1 * x_main + beta_2 * x_2 + beta_3 * x_3 + beta_4 * x_4 + rnorm(n, 0, sigma)
  pvals[i,] = summary(lm(y ~ x_main + x_2 + x_3 + x_4))$coefficients[2:5, 4]
}

out_data = melt(pvals)[,2:3]
colnames(out_data) <- c("Coefficient", "pvalue")

ggplot(out_data, aes(x = log(pvalue), fill = as.factor(Coefficient))) +
  geom_histogram(alpha=0.5, position="identity", bins = 30) +
  theme_bw()
```




```
library(reshape2)

nsims = 1000
n = 20
beta_1 = 2
beta_2 = 2
beta_3 = 2
beta_4 = 2
sigma = 2
```

```r
# Model with X_1 as the only predictor
pvals = matrix(nrow = nsims, ncol = 4)
for (i in 1:nsims) {
  x_main = rnorm(n, 0, 2)
  x_2 = rnorm(n, x_main, 2)
  x_3 = rnorm(n, x_main, 2)
  x_4 = rnorm(n, x_main, 2)
  y = beta_1 * x_main + beta_2 * x_2 + beta_3 * x_3 + beta_4 * x_4 + rnorm(n, 0, sigma)
  pvals[i,1] = summary(lm(y ~ x_main))$coefficients[2, 4]
  pvals[i,2] = summary(lm(y ~ x_2))$coefficients[2, 4]
  pvals[i,3] = summary(lm(y ~ x_3))$coefficients[2, 4]
  pvals[i,4] = summary(lm(y ~ x_4))$coefficients[2, 4]
}

out_data = melt(pvals)[,2:3]
colnames(out_data) <- c("Coefficient", "pvalue")

ggplot(out_data, aes(x = log(pvalue), fill = as.factor(Coefficient))) +
  geom_histogram(alpha=0.5, position="identity", bins = 30) +
  theme_bw()
```



In the multiple regression model, β_1 is usually much less significant than the auxiliary predictors even though it is the main predictor. However, in the single regression model, it is much more significant. By fiddling with the coefficients in the multiple regression model, we can see that increasing the β coefficient makes a predictor more likely to be significant, and increasing the variance of an $\vec{X}$ increases the significance of its β coefficient. In general, increasing the variance of $\vec{X}_i$ increases the significance of its coefficient: in the original

setup, $\vec{X}_i$ for $i \neq 1$ had higher variances than $\vec{X}_1$, so they had lower p-values.